MEASUREMENT IN SECONDARY EDUCATION

THE MACMILLAN COMPANY
NEW YORK · BOSTON · CHICAGO · DALLAS
ATLANTA · SAN FRANCISCO

MACMILLAN AND CO., LIMITED
LONDON · BOMBAY · CALCUTTA · MADRAS
MELBOURNE

THE MACMILLAN COMPANY
OF CANADA, LIMITED
TORONTO

MEASUREMENT
IN
SECONDARY EDUCATION

BY

PERCIVAL M. SYMONDS

ASSOCIATE PROFESSOR OF EDUCATION
TEACHERS COLLEGE, COLUMBIA UNIVERSITY

New York
THE MACMILLAN COMPANY
1938

COPYRIGHT, 1927,

BY THE MACMILLAN COMPANY.

All rights reserved — no part of this book may be reproduced in any form without permission in writing from the publisher, except by a reviewer who wishes to quote brief passages in connection with a review written for inclusion in magazine or newspaper.

Set up and electrotyped. Published July, 1927. Reprinted July, 1928, twice; February, July, 1929; June, August, 1930; September, 1934; May, 1938.

• PRINTED IN THE UNITED STATES OF AMERICA •

PREFACE

It is now nearly two decades since the movement to measure educational products started. The early work was experimental and to a large degree theoretical. The first applications of the measurement movement were made in the elementary field. It was not until after the war that the newer methods of objective measurement were generally used in the construction of examinations on the secondary and higher levels and standardized tests in high school subjects began to appear. Since 1920 the construction and use of objective and standardized tests in the high school has gathered momentum until at present there is a sizable body of material and of techniques available. It is the purpose of this volume to bring together these results, which are to be found mainly in technical periodicals, and make them more generally available.

This book is addressed first of all to high school teachers and principals, and secondarily to students of measurement. It aims to be an introductory book. Its order has been determined by the experience which the author has had in conducting a course bearing the same title as this book. A limited number of standardized tests have been described somewhat at length. Several of them will undoubtedly be supplanted soon by better ones. The purpose of the chapters in which standardized tests are described is not only to introduce them to the reader but to make the discussion of them bring out in a concrete way the characteristics of a good test.

Measurement in secondary education makes use of certain facts and principles which are different from those in the older

and better established measurements in the elementary school. (1) The subject matter of the high school has a natural unit of a year (science, mathematics, or history) or three or four years (foreign language) as against several years (arithmetic, reading, spelling, language) in the elementary school. (2) There is a change in the rate of mental growth during the high school years the nature of which is only partly understood. (3) In secondary education the unit is the subject, whereas in elementary education the unit may be the school year. (4) In the junior high school differentiation begins, and throughout secondary education there is an increasing opportunity for choice. These conditions raise specific problems in connection with the use of derived units in interpreting test scores, in the grouping of pupils, in marking, and in prediction and guidance.

There is a tendency for those interested in measurements in high school to divide into two camps: those who are interested in the development of the informal new-type examination and those interested in the development of standardized tests. This book takes a middle ground. Its thesis is that a truly scientific use of measurements in the high school awaits the development of standardized tests in the high school subjects. But it recognizes the difficulties and dangers that accompany the development of the standardized test. Informal tests are recognized as important and essential instruments which must be used until they can be supplanted by standardized tests.

I take this opportunity to pay tribute to my teachers of measurement. I sat under Dr. M. R. Trabue in a course which is the direct ancestor of this book. Dr. E. L. Thorndike and Dr. W. A. McCall, pioneers in the development of educational measurements, have shared in teaching me, both in class and by their writings, the fundamental principles of measurements. Dr. T. L. Kelley has been an inspiration and guide in the use of statistics. I also wish to acknowledge my indebtedness at this time to those

who have done more than others to further the development of measurements in secondary education, particularly Dr. B. D. Wood, Dr. G. M. Ruch, Dr. H. A. Toops, and Dr. A. S. Otis. Many others have contributed to the movement and their names may be found in the chapter bibliographies. I wish also to acknowledge the counsel and help of my wife, Johnnie Pirkle Symonds.

Percival M. Symonds

CONTENTS

CHAPTER		PAGE
I.	Why Measurement in High School?	1
II.	Why Better Measurement in High School?	10
III.	Methods of Improving Examinations	24
IV.	Intelligence Tests for Use in High School	53
V.	Standardized Tests for the Teacher of English	84
VI.	Standardized Tests for the Teacher of Mathematics	113
VII.	Standardized Tests for the Teacher of Science	130
VIII.	Standardized Tests for the Teacher of Language	148
IX.	Standardized Tests for the Teacher of Social Science	162
X.	Tests for Physical Education	184
XI.	Miscellaneous Tests	206
XII.	How to Handle the Results of Testing	231
XIII.	Norms and Standards	254
XIV.	Criteria for the Choice of Tests	278
XV.	Derived Scores	310
XVI.	The Measurement of Conduct	333
XVII.	Prognosis Tests	363
XVIII.	Tests of Clerical and Mechanical Ability	377
XIX.	Prediction of Success in High School	390
XX.	Prediction of Success in College	411
XXI.	The Use of Tests in Guidance	448
XXII.	Promotion	465
XXIII.	Ability Grouping	474
XXIV.	Marks and Marking Systems	498
XXV.	The Use of Informal Tests in Teaching	531

CHAPTER PAGE

APPENDIX

A. STATISTICAL COMPUTATIONS 543

B. TECHNICAL PROCEDURES IN DETERMINING TEST RELIABILITY 561

C. LIST OF PUBLISHERS WHO PRINT TESTS . 568

INDEX . . . 569

LIST OF TABLES

TABLE		PAGE
1.	Passing points in 206 Illinois high schools	4
2.	Original marks and marks assigned by ninety-one teachers who regraded three answers to a question in geography	12
3.	Percentage of failure in algebra on College Entrance Board Examinations from 1916 to 1921	14
4.	Distribution of marks in the University of Chicago High School for 1908–1909 (from Johnson)	17
5.	Otis Group Intelligence Scale, Higher Examination: Tests, time allowance and possible score	59
6.	Otis Group Intelligence Scale, Advanced Examination: Grade norms	63
7.	Otis Group Intelligence Scale, Advanced Examination: Age norms	63
8.	Otis Group Intelligence Scale, Advanced Examination: Mental age norms	64
9.	Otis Self-Administering Test of Mental Ability, Higher Examination: Grade norms	66
10.	Otis Self-Administering Test of Mental Ability, Higher Examination: Age norms	66
11.	Difference between mean scores, Alpha forms	67
12.	Army Alpha: Test, time allowance, and possible score	67
13.	Army Alpha: Grade norms	68
14.	Army Alpha: Mental age norms	68
15.	Terman Group Test of Mental Ability: Tests, time allowance, and possible score	69
16.	Terman Group Test of Mental Ability: Percentile norms by grade	70
17.	Terman Group Test of Mental Ability: Mental age equivalents	71
18.	Miller Mental Ability Test: Tests, time allowance, and possible score	72
19.	Miller Mental Ability Test: Grade norms	72
20.	Miller Mental Ability Test: Percentile grade norms	73
21.	Thorndike Intelligence Examination for High School Graduates: Tests and time allowance	75
22.	Thorndike Intelligence Examination for High School Graduates: Frequencies of scores in per cents	75
23.	Thurstone Psychological Examination for High School Graduates and High School Seniors, No. IV: Norms	76

TABLE		PAGE
24.	Scholastic Aptitude Test: Tests, time allowance, and number of items	77
25.	Hudelson Typical Composition Ability Scale: Grade norms (January)	88
26.	Showing computation for determining average error and direction of error in rating a composition by means of a composition scale	89
27.	Sixteen Spelling Scales: February norms given in terms of per cent	96
28.	Pressey Diagnostic English Composition Tests: Grade norms	103
29.	Haggerty Reading Examination, Sigma 3: Tests, method of scoring, possible score, and time allowance	107
30.	Haggerty Reading Examination, Sigma 3: Grade norms	107
31.	Hotz Algebra Scales: Norms	116
32.	Rugg-Clark Algebra Tests: Tests, time allowance, and illustrative examples	121
33.	Rugg-Clark Algebra Tests: Norms	122
34.	Schorling-Sanford Geometry Test: Tests, possible score, and time allowance	128
35.	Schorling-Sanford Geometry Test: Norms (June)	128
36.	Ruch-Popenoe General Science Test: Tests, time allowance, and possible score	132
37.	Ruch-Popenoe General Science Test: Norms (June)	133
38.	Ruch-Cossman Biology Test: Tests, number of items, and time allowance	134
39.	Ruch-Cossman Biology Test: Tentative percentile norms	135
40.	Iowa Physics Test: Tests, number of items, and time allowance	137
41.	Iowa Physics Test: Percentile norms	137
42.	Powers General Chemistry Test: Percentile norms	140
43.	Glenn-Oburn Physics Tests: Tests, time allowance, and number of questions	143
44.	Glenn-Welton Chemistry Tests: Tests, time allowance, and number of questions	144
45.	Ullman-Kirby Latin Comprehension Test: Semester norms	152
46.	White Latin Test: Semester norms	153
47.	Godsey, Pressey Syntax, and Tyler-Pressey Latin Tests: Norms	156
48.	American Council Modern Language Tests: Norms	159
49.	Percentage of concurrence of teachers of history in aims listed in a questionnaire	163
50.	Van Wagenen American History Scales: "Uncorrected score"	168
51.	Barr Diagnostic Tests in American History: Norms	173
52.	Pressey-Richards Test in American History: Tests, time allowance, and October norms	175
53.	Brown-Woody Civics Test: Percentile norms (end of year)	181
54.	Schneider Index: Points for grading cardio-vascular changes	190
55.	Weight-Height-Age table for boys of school age	193

TABLE PAGE

56. Weight-Height-Age table for girls of school age 194
57. Chest capacity: Norms in cubic centimeters (from Smedley) . . 196
58. Vital Index: Norms (from Smedley) 196
59. Grip: Norms in kilograms (from Smedley) 197
60. Kwalwasser-Ruch Test of Musical Accomplishment: Tests, maximum score, and time allowance 212
61. Kwalwasser-Ruch Test of Musical Accomplishment: Grade norms and percentile norms for the high school 213
62. Home Economics Information Tests: Tests and number of items . 218
63. Murdoch Sewing Scale: Norms 220
64. Blackstone Typewriting Test: Norms 228
65. Illustrating the construction of a frequency distribution . . . 233
66. Illustrating two distributions having the same mean but different variabilities 236
67. Distribution of scores on the Terman Group Test of Mental Ability in a private school 243
68. Illustrating a correlation of 0.00 248
69. Illustrating a correlation of 0.25 248
70. Illustrating a correlation of 0.50 249
71. Illustrating a correlation of 0.75 249
72. Illustrating a correlation of 1.00 250
73. Averages on a spelling test given in thirteen fourth-grade rooms in a city school system 255
74. Average scores made by 12-year olds in city and country schools (from Frost) 266
75. Median scores made by white recruits in Army Alpha in forty-one states (from Alexander) 267
76. Mean score in intelligence on a combined scale derived from Army Alpha and Army Beta of foreign-born white recruits (from Brigham) . . 269
77. Van Wagenen American History Scales: Grade medians . . . 270
78. Speed of reading in words per minute on Courtis Silent Reading Test (from O'Brien) 271
79. Comprehension of reading as determined by number of questions correctly answered in Courtis Silent Reading Test 272
80. Number of individuals making extreme plus or minus deviations on repetitions of the National Intelligence Tests 294
81. Reliability coefficients reported by Wood, Toops, and Ruch for new-type tests 297
82. Otis Self-Administering Test of Mental Ability, Norms intermediate examination 315
83. Distribution of mental ages and I Q's in the graduating elementary school class, Providence, R. I. (adapted from Allen) . . . 323

TABLE PAGE

84. Distribution of Arithmetic AR derived from the National Intelligence Scale *B* and the Woody-McCall Arithmetic tests 325
85. Distribution showing number of classes in fifty-four rating scales (from Boyce) 340
86. The amount of disagreement among judges in estimating, on the basis of acquaintance, the traits of others, in two investigations (from Hollingsworth) 351
87. Number of ratings necessary to obtain a specified reliability for various traits 353
88. Correlations of tests with criterion of mathematical ability (from Rogers) 367
89. Correlation of Rogers Test of Mathematical Ability with school marks 368
90. Intercorrelation of algebraic ability, geometry ability, and verbal ability (from Rogers) 369
91. Rogers Test of Mathematical Ability: Tests and time allowance . . 370
92. Correlation of tests with criterion of Latin ability (from Allen) . . 371
93. Correlations of Allen Latin Prognosis Tests 371
94. Allen Latin Prognosis Test: Tests, time allowance, and correlation with criterion 372
95. Correlations of National Intelligence Test with average of four scholastic tests 375
96. Correlations of Stanford-Binet with average of four scholastic tests . 375
97. Stenquist Mechanical Assembly Test: Age norms 385
98. Correlation of first-year average grade in high school with various factors (from Kelley) 393
99. Correlation of marks in first year of junior high school with various tests (from Fretwell) 394
100. Average correlations between marks in the elementary school and marks in the first year of high school (from Ross) 395
101. Correlation of marks in the junior and senior high schools with various factors (from Flemming) 397
102. Correlation between first-year high school marks and intelligence tests (from Proctor) 397
103. Correlation of high school marks and group intelligence tests (from Jordan) 398
104. Average correlations of high school marks and group intelligence tests 398
105. Résumé of correlations between high school marks and other factors . 399
106. Correlation of a criterion of Latin ability with various factors (from Clem) 403
107. Correlation of group intelligence tests with separate high school subjects 405
108. Correlation of college marks with secondary school marks . . . 413
109. Correlations between college marks and secondary school marks, University of Hawaii 415
110. Correlation of college marks and subject matter examinations . . 416

TABLE		PAGE
111.	Average correlation of various intelligence tests with college marks	419
112.	Comparison of combined high school and intelligence ratings with freshman scholarship (from Johnston)	424
113.	Coefficients of correlation between high school marks and college marks for separate subjects	430
114.	Correlation between college marks and subject-matter examinations and school marks for separate subjects (from Whitman)	432
115.	Correlation of college marks and Iowa examinations	433
116.	Correlations of college marks and Iowa Placement Examinations in the University of Arkansas (from Stoddard)	434
117.	Curricula offered in high school in fifteen American cities — Adapted (from Counts)	449
118.	Occupational intelligence standards based on Army Alpha Intelligence Tests (from Fryer)	455
119.	Distribution of French Vocabulary Test (American Council Test)	471
120.	Ruch-Popenoe General Science Test: Distribution of Scores (June)	476
121.	Shifting after re-grouping — Indicating permanency of placement in ability grouping	479
122.	Marking systems in use in 281 Illinois high schools	503
123.	Average scores on a standardized French test and average mark given by teachers in eight sections of senior French in a city high school	508
124.	Suggested form for the study of a school's distribution of marks	509
125.	Percentage distributions for classes of different sizes and for scales having different number of points	512
126.	Scores and suggested marks for a class of thirty-six pupils in general science	513
127.	Studiousness index, using the method of ranks	522
128.	Studiousness index, using standard deviation units	523
129.	Time table for constructing new-type tests for high school or college students and adults (from Ruch)	533
130.	Showing method of determining goodness of test item	536
131.	Chronological ages and scores on the Terman Group Test of Mental Ability in two classes of a secondary school	546
132.	Scatter-diagram: Scores on the Hotz Equation — Formula Test and Scores on the Haggerty Reading Sigma 3	551
133.	The Computation of the Pearson Coefficient of Correlation	552

LIST OF FIGURES

FIGURE		PAGE
1.	Posture Scale	188
2.	Distribution of frequencies given in Table 66	239
3.	Distribution of frequencies given in Table 66	240
4.	Frequency polygon of data given in Table 66	241
5.	Normal probability curve designed to fit data of Table 66	241
6.	Bi-modal frequency curve	241
7.	Normal probability curve	242
8.	Curve of mental growth	312
9.	Curves of mental growth of average and superior children	313
10.	Curves of mental growth from age norms and mental age norms	315
11.	Distribution of I Q's	317
12.	Distribution of I Q's for 905 unselected children 5–14 years of age (from Terman)	318
13.	Profile graph suggested for the guidance of pupils in high school	459
14.	The growth of a school group in some subjects	467
15.	Two methods of grouping pupils	475
16.	Three methods of grouping 250 pupils into 10 sections	489
17.	Progress of an average (*B*) section, and three fast-moving sections (A_1, A_2, and A_3) starting at three different times	492
18.	Showing progress on nine standardized tests throughout a school year: a suggested form for recording absolute marks	501
19.	The normal distribution of marks	514
20.	Two methods of marking fast- and slow-moving sections	525
21.	Distribution of means	557

MEASUREMENT IN SECONDARY EDUCATION

CHAPTER I

WHY MEASUREMENT IN HIGH SCHOOL?

The high school has always had measurement and the current interest in the measurement movement is an attempt to make it better. Every judgment made of a student is a measurement. The teacher who reads and grades a number of English compositions is performing a kind of measurement. Perhaps she sorts the papers into five piles according to merit, and roughly makes for herself a five-point scale. If, on the basis of these compositions, this teacher forms opinions as to the general abilities of her pupils, she is, in a rough way, measuring intelligence. Likewise, if a pupil is given a failing mark in algebra or chemistry, a very real and very valuable form of guidance, both as to educational and vocational possibilities, is being offered, for the pupil is discouraged in a decisive way from following a scientific career or even from pursuing further his education along scientific lines. Judgments of pupils' achievement or intelligence or personal qualities are being made constantly by teachers. The present movement in educational measurements endeavors to make these judgments more exact.

Measurements are being used now in high school for the following purposes:

1. To inform pupils of their achievement.
2. As incentives to study.

3. To promote competition
   a. Between groups;
   b. Between individuals;
   c. With one's past record.
4. To determine promotion.
5. To diagnose weak spots in the pupils' achievement.
6. To determine the quality of instruction.
7. To determine admission to high school.
8. To place a pupil in the school.
9. To determine admission to college.
10. To provide reports to parents.
11. To determine credits, honors, etc.
12. Educational and vocational guidance.
13. To rate teachers.
14. To predict a pupil's success.
15. To study the efficiency of the school.

Measurements are being made in all schools, either in the form of offhand judgments, or with the crude instrument which I designate as the traditional examination; and the results are being expressed in terms of the percentage marking system or the letter system so well known in high school practice. Measurements are not an innovation in high school. The purpose of this book is not to invent measurements, but to show how to make them better.

The balance of this chapter describes the nature of measurements in high schools as they have come down to us in tradition. Measurements of this type were universal in high schools ten years ago and are still found in high schools to-day which have remained immune to the contagion of new ideas. The next chapter presents certain criticisms of these traditional methods of measurements and the remainder of the book is devoted to a description of newer measuring instruments and how they may be most effectively used.

The basis of measurements has been the traditional examination. I assume all who read this are familiar with this instrument, as we have all taken these examinations many times in our own school careers. For the sake of definiteness I include samples.

Examination in Biology

1. Define (*a*) chlorophyll, (*b*) protozoa, (*c*) protoplasm, (*d*) osmosis.
2. Trace the life history of mold.
3. What are the distinguishing characteristics of moths and butterflies?
4. Draw a diagram of a complete flower and label.
5. Describe the adaptation of flowers for preventing self-pollination.
6. Name three orders of insects, with two examples of each.
7. What is the theory of evolution? Who is the originator?
8. State how the following measures aid in controlling the spread of epidemics: (*a*) quarantine, (*b*) vaccination, (*c*) boiling drinking water, (*d*) school nurses.
9. What are the functions of leaves?
10. Name two foods rich in protein, grape sugar, and starch.

Examination in United States History

1. Define: (*a*) Squatter Sovereignty, (*b*) the Missouri Compromise, (*c*) the Monroe Doctrine, (*d*) the Pendleton Act, (*e*) the Virginia-Kentucky resolutions, (*f*) the Spoils System.
2. Name the five most famous Spanish explorers and their explorations.
3. What were the conditions in Europe that led to the discovery of America?
4. Why did the United States enter the World War?
5. How can the Constitution of the United States be amended?
6. Tell who each of the following was and name the thing he did: (*a*) Henry Clay, (*b*) John Marshall, (*c*) Alexander Hamilton, (*d*) Benjamin Franklin, (*e*) Daniel Webster, (*f*) Thomas Jefferson, (*g*) William L. Garrison, (*h*) Calhoun, (*i*) G. R. Clark, (*j*) Roger Williams.
7. What motives led immigrants to America?
8. Give the provisions of the Compromise of 1850.

These tests are readily made by the teacher to fit the subject matter which has been covered in the foregoing term. Not much

attention is given to the wording of the questions or to the precise type of mental operation involved. Pure memory questions are frowned upon by progressive teachers although inspection of actual practice will show that they are still very common.

These tests are usually scored on a percentage basis. For a perfect paper the score is 100. For a paper in which nothing is correct the score is zero. Usually questions are assigned arbitrary credits before the examination is given. This defines the maximum per cent that can be made on any single question. These credits are usually intended to bear some relation to the difficulty of the question. The questions hardest to answer usually receive the most credit, although I have heard of a system of scoring where the opposite practice is followed and the hardest questions are given the least credit. A pupil's mark on the examination is then the sum of the credits on each separate question. Practically all schools have an arbitrary passing point in terms of per cent. The following table from Odell[1] shows the passing points in 206 high schools in Illinois using the percentage system:

TABLE 1

PASSING POINTS IN 206 ILLINOIS HIGH SCHOOLS

	NUMBER
PASS AT 70	
No conditions	40
Conditions at 60, 65, 67, 68, 69 or indefinite	16
PASS AT 75	
No conditions	81
Condition at 70	38
Condition at 60, 65, 72, 73, 74 or indefinite	30
PASS AT 80	1
	206

[1] ODELL, C. W. — "High School Marking Systems," *School Review*, 33 : 346–354.

There is a tacit assumption that a school with a higher passing point is a school with higher standards. A school with a passing point of 80% is credited with having considerably higher standards than a school with a passing point of 60%. With this examination system we have all become per cent-minded with reference to school work.

The results of examinations are summarized in school marks given at the end of certain periods, as at the end of semesters, terms, or even months, to indicate the quality of work done during the period. These marks are supposed to include the results of examinations and other evidences of the pupils' ability as accumulated by the teacher in the daily class period. Some schools go so far as to determine that the final term mark shall be based one third on the final examination, one third on informal quizzes given throughout the term, and one third on class work. Several letter systems are in vogue. One of the most common is allotting *A*, *B*, *C*, and *D* to four degrees of quality of passing work and *F* to work of a failing quality. These letter grades are generally preferred to per cents, since there are fewer letters than per cents and such fine distinctions are not demanded. That is, in the system above outlined there are only four passing letters, *A*, *B*, *C*, and *D*, whereas there are forty percentage steps from 60 to 100. Some teachers who wish to make finer distinctions than *A*, *B*, *C*, *D*, and *F*, use plus and minus; thus *C*+ means better than *C*, and *B*− means better than *C*+ but not as good as *B*.

This is about the extent of measurements in the old-fashioned high school. Pupils are informed of their achievement by the percentages on examinations and quizzes. These examination marks are the only incentives to study that the pupils gain as a result of their work — all other incentives must come as pressure from without. Term marks also inform pupils of their school standing. They are usually considered as rewards or prizes to be wrung from the teacher rather than as the result of the pupil's

growth in learning. This last point is very important. In the old-fashioned school it is very seldom that the pupil is made to consider himself a learner. Hence under this system his effort is directed toward impressing the teacher rather than in mastering subject matter.

Under this system very little individual diagnostic work is done. When a pupil fails, no analysis is made of the fact. A failure is a failure and can be made up only by repeating the course or being tutored. The work made up is tested by an examination covering the work of the course even though the pupil has shown himself deficient in only one phase of the work.

Pupils are admitted to high school on the basis of similar marks in the elementary school. Colleges in the east, however, have not been willing to accept high school marks at their face value and have set their own examinations or have made use of the examinations set by the College Entrance Examination Board. In the west a more liberal attitude has established the system whereby the marks of an accredited school are accepted as a means of determining admission to college. Promotions within the high school have been determined entirely on the basis of marks. As stated before, each school sets up its own passing standard in terms of its marking system and that passing mark is made the dividing point between promotion or repetition of the subject or class.

Reports to parents are again in terms of these same marks. As most parents are already familiar with the system in their own school experience, report cards are readily interpreted. The critical point is the passing point, as that tells whether or not the child will progress with the rest of his classmates. Failure implies failure of promotion and that means "something must be done." Special credits — honors, prizes, and the like — are also determined from these same marks. One school that the writer knows stars the names of pupils on the graduation program who

make an average mark of *B* or over, thus creating an honor rank. *A* and *B* have come to be considered as honor marks, and ambitious parents are not satisfied unless their children attain *A* or *B* in their studies. Less ambitious sons and daughters, however, are only too satisfied with a *C*; and at some colleges a mark of *C* used to be known as a "gentleman's mark," because it could be achieved while living the life of a "gentleman."

The efficiency of instruction is gauged by the number of failures in a school. If a teacher has a large percentage of failures, her efficiency is at once questioned. If the percentage of failures is low, the teacher is a successful teacher. For the teacher, it is always a constant battle between the examination results and the judgment that will be pronounced by the authorities when the marks are reported. A young and inexperienced teacher will set stiff examinations, abide by the results, and suffer the consequences in failures and in being judged as an unsuccessful teacher. With experience the same teacher will learn the secret of reducing the number of failures and gaining the consequent approval of pupils and administration by setting easier examinations. This latter procedure is not possible, however, in schools which prepare for college examinations, as duplicity on the part of the teacher in marking will be later exhibited in the results of the college entrance examinations.

With these applications marks have been stretched almost to the limit. In the older schools there is practically no attempt at ability sectioning. When the school is large enough to have more than one section in a subject, strange methods of sectioning are employed. Pupils are sectioned according to the alphabetical order of their names, or in the order in which they were registered, or according to the exigencies of program making. But often failure to section according to ability is due to a definite belief that such a procedure is harmful and undemocratic. Practically no use of marks is made for prediction or prognosis. The main

reason for this is a firm belief that marks do not predict. "Anything is possible" was the belief of the older educator. Any boy is a possible "A." One hundred per cent is a possible goal for all. Likewise failure stares every pupil in the face and fear of failure is the main incentive for striving. When pupils fail to achieve the maximum, it is moral perversity or obstinacy or laziness or some other moral defect. "If you would only quit your fooling and get down to work you could pass, I am sure" is the teacher's attitude. With such a philosophy, prediction is an impossibility. It is like trying to predict the temperature to-morrow or the size of a cloud in half an hour.

With prediction an impossibility, guidance is not attempted. That high school is to prepare for college is the guiding principle, and so the school is organized around the college preparatory course. For those who do not wish to go to college a general course is provided. A commercial course cares for those whose interests are still less academic. If interest lags and enthusiasm for Latin and algebra wanes, it is best to get out and go to work. This again is a moral perversity, a kind of wildness that makes a boy spurn the advantages offered to him and wilfully refuse them. Some pupils, who are persuaded to believe that all virtue lies in the path to college, manfully stick to the course even with failure piled on failure — confident that industry can win.

The efficiency of the school is also a new conception in education. Previously the principal's only measure of his school's efficiency was the number of boys and girls he could get into college, their success with the College Entrance Examination Board's examinations, and their success in college as compared with boys from other schools. Direct comparison with other schools was not possible because the technique of this comparison was not available.

Why measurements in secondary education? Because measurement is essential in the education process. Psychologists tell us that learning takes place more readily when the results are

accompanied by satisfaction. Much of the learning in school does not result in immediate satisfaction as does the hitting of a target or the drop kicking of a ball between the goal posts. In fact, much of the learning in school is like driving a golf ball over a hill. One may practice indefinitely driving a golf ball over a hill without any gain in accuracy or distance, because he cannot see where the ball lands. Activity is essential in learning, but one must be able to measure success or failure and improvement. That is what measurement in education aims to accomplish.

REFERENCES

JOHNSON, F. W. — *Administration and Supervision of the High School*, Ginn & Company, Boston, 1925.

ODELL, C. W. — "High School Marking Systems," *School Review*, 33: 346–354, May, 1925.

THORNDIKE, E. L. — "The Nature, Purposes and General Methods of Measurements of Educational Products." THE SEVENTEENTH YEARBOOK OF THE NATIONAL SOCIETY FOR THE STUDY OF EDUCATION, PART 2, *The Measurement of Educational Products*. Public School Publishing Company, Bloomington, Ill., 1918.

CHAPTER II

WHY BETTER MEASUREMENT IN HIGH SCHOOL?

Since measurement is a necessary part of the educative process, it is worth while trying to improve it and to make it as accurate as possible. It is often surprising for a teacher to learn that her judgments of pupils are much in error or differ widely from those of other persons. Better measurements are needed in high school because the present measurements are faulty.

In the first place, the scoring of examination papers is an exceedingly subjective process. Ruch [1] has demonstrated this by an experiment. To ninety-one teachers he submitted a single question in geography, together with three answers actually written by pupils in a geography class. These answers were selected from the best, median, and poorest papers in the class. The question, directions for grading it, and the three answers are repeated below:

Directions: Below are three actual answers to the question: "Name and locate five of the largest cities of the United States and name their leading industries, exports, and imports."

Grade each of the three answers on a scale of 0 to 20, according to your best judgment of its merit, 20 being an answer ordinarily accepted by teachers as entirely satisfactory, and 0 being an answer practically without discernible merit.

Answer 1

Five of the largest cities in the United States is Detroite. An export is Cars. And industry is Manufactoring. Chicago is an important city and an export is Manufactored and canned goods. An industry of Chicago is

[1] Ruch, G. M. — *Improvement of the Written Examination*, pp. 54–57, 1924.

meate packing. New York is another important city. An industry of N. Y. is manufactoring. An export of N. Y. is manufactored goods. Pittsburge is an important city of U. S. An export is iron core. A industry of Pittsburge is manufactoring. Another important city of U. S. is New Orleans. An export of New Orleans is cotton. An industry of New Orleans is manufactoring.

Grade...............................

Answer 2

The five largest cities of the United States are (1) navada (2) Arkansas. The leading industry of navade is manufacture and the leading industry of Arkansas is agriculture. The leading imports are manufacturing mostly.

Grade...............................

Answer 3

The 5 largest cities of United States are New York, Chicago, St. Louis, Boston, San Francisco. New York is in the State of New York. Chicago is in the state of Illinois. St. Louis is in the State of Missourri. Bostin is in the state of Mass. San Fransisco is in the state of California. New York is a manufacturing city. Chicago is noted for meat packing center. St. Louis is noted for manufacturing textile goods and iron goods. San Fransisco is noted for the packing of fruit. New York exports iron goods and imports wool, cotton, and other raw materials. Chicago exports meat and hides and grain and imports food and grains. St. Louis exports manufactured products and imports raw materials.

Grade...............................

The evidence here is decisive. In such a simple act as judging the value of answer 1 on a scale of 20, one teacher assigned a value of 2 and another a value of 20. That the experience of these teachers in grading these answers is typical of the marking of examinations in general is evidenced by the investigations by Starch and Elliott.[1] These investigators selected a final ex-

[1] STARCH, D. and ELLIOTT, E. C. — "Reliability of Grading High School Work in Mathematics," *School Review*, 21 : 254–259, April, 1913. — "Reliability of Grading High School Work in English," *School Review*, 20 : 442–457, Sept., 1912. — "Reliability of Grading High School Work in History," *School Review*, 21 : 676–681, Dec., 1913.

TABLE 2

ORIGINAL MARKS AND MARKS ASSIGNED BY 91 TEACHERS WHO REGRADED THREE ANSWERS TO A QUESTION IN GEOGRAPHY

Mark	Answer 1	Answer 2	Answer 3
20	1		9
19			3
18	1		21
17	1		12
16	2		17
15	10		15
14	2		3
13	8		1
12	14		3
11	5		1
10	24		3
9	5		1
8	7		1
7	3		
6	4		
5	1		
4			
3	2		1
2	1	3	
1		1	
0		87	
Number	91	91	91
Mean	10.9	0.1	16.1
Standard Deviation . .	3.2	0.4	2.9
Original Mark	16	2	12

amination paper in geometry, written by a student in one of the largest high schools in Wisconsin. An exact reproduction of this paper and a set of questions were sent to one hundred and eighty high schools in the North Central Association. It was requested that this paper be graded according to the practice and standards of the school by the principal teacher of mathematics. One hundred and sixteen acceptable replies were received. The paper

showed evidence of having been marked with unusual care and attention. In seventy-three schools where the passing grade was 75, the lowest mark given was 39 and the highest 88. The mode was 75, with twelve teachers giving this mark. Of the one hundred and sixteen marks assigned to this paper, two were above 90 and one below 30. Twenty were 80 or over and twenty other marks were below 60. Forty-seven teachers assigned a mark passing or above, but sixty-nine thought it not worthy of a passing mark. Similar results were obtained in scoring an examination in American history and first-year high school English.

Wood[1] tells of inadequacies of marks given in correcting college entrance examinations in algebra and geometry in June, 1921. About 400 algebra papers and an equal number of geometry papers selected at random were each scored twice independently by two different readers of the board. The scoring was such that agreement on failures would be somewhat as follows: in gross numbers suppose that 10,000 candidates are tested with form *A* of a given geometry examination and that 30% of the 10,000 fail. The same 10,000 are tested with another equivalent geometry examination, say form *B*, which also fails 30% of the candidates.

Failed Form *A* Passed Form *B* 1,279	Passed on both forms 5,721
Failed on both forms 1,721	Passed Form *A* Failed Form *B* 1,279

In other words, accepting the results of one such examination as valid, which was done by the Board in 1921, another equivalent examination would pass 1,279 of the 3,000 failed by the first, and would fail 1,279 of the

[1] Wood, B. D. — *Measurement in Higher Education*, pp. 124–125. Copyright 1923 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

7,000 passed on into college by the first. The fate of over 2,500 in 10,000 candidates would be reversed by an equivalent form of the same examination when the reliability is no higher than that of the *C. E. E. B. Mathematics C Examination* for June, 1921.

Even the supposition of equivalent examinations is more than can be expected from successive College Entrance Examination Board examinations. This introduces another large and unsuspected source of error. Wood says:[1]

Large and unrecognized differences were found between the difficulties of the examinations which were thought to be equal. According to the reports of the Secretary of the College Entrance Examination Board the percentages of failures in algebra for the years 1916 to 1921 inclusive have been as follows:

TABLE 3

PERCENTAGES OF FAILURE IN ALGEBRA ON COLLEGE ENTRANCE BOARD EXAMINATIONS FROM 1916 TO 1921

Year	1916	1917	1918	1919	1920	1921
Percentages of failures	61.8	36.7	25.3	61.3	26.1	28.5

These variations are almost certainly due to differences in the difficulties of the examinations used.

A humorous incident illustrating the same point is told by Wood[2]:

The facts of a subjective scale are well illustrated in the following anecdote concerning the grading of history papers by a group of college professors of history in the summer of 1920. One of the five or six expert readers assigned to a certain group of history papers, after scoring a few, wrote out for his own convenience what he considered a model paper for the given set of ten questions. By some mischance this model fell into the hands of another reader who graded it in a perfectly bona fide fashion. The mark he assigned to it was below passing, and, in accordance with the

[1] *Op. cit.*, p. 128 f.

[2] WOOD, B. D. — "Measurement of College Work," *Educational Administration and Supervision*, 7: 301–304, Sept., 1921.

custom, this model was rated by a number of other expert readers in order to insure that it was properly marked. The marks assigned to it by these readers varied from 40 to 90.

These results all point in the same direction — the marking of examination papers is highly subjective. Ruch [1] says as a result of his investigation:

> It seems more advisable to attempt improvement by changing the type of the examination than by refining the traditional methods. Human judgments are fallible at best, and improvement in marking essay-type examination papers can only take place within fairly narrow limits.

Traditional examinations are corrected on a percentage basis. If ten problems in arithmetic are given and eight of them are right, the test is marked 80%. This method of marking has so invaded the schools of the country that practically every teacher tends to picture students in terms of per cents. This is a 90% boy, this is a 60% boy. By a trial and error process it was determined that if a boy could do on the average 60% or 70% of the work assigned him, he usually could manage the work of the next grade or year. So the whole theory of promotion and placement in school has pivoted around this percentage scale which labels children according to the percentage they attain on tests set them.

There are several fallacies in this system. First, 100% means perfection. Of course 100% means perfection on the test given, but the test given in no case represents the ultimate in any subject. There is always more to be learned in any field and questions may always be more complex or difficult than any on the test in which 100% stands for perfect. It is also possible to learn to answer the questions on the test in less time. Perfection is unattainable in any phase of school work and one of life's greatest delusions is the delusion of 100% in school. It is an illusory goal to work for and it is a delusion and a snare when attained. It is the antithesis of growth and progress.

[1] *Op. cit.*, p. 62.

A second fallacy in the percentage system is that a given per cent — say 60% — is passing. All depends upon the test set. Make your test easy and many pass — make your test hard and many fail. It is not merely a question of testing the material that has been taught. Making a test long within the time allowed, or making the wording of the questions complex will of itself make a test hard. Passing on a basis of 60% means passing on a variable standard. Then again, shall the passing mark be 60% or 70% or even 75%? Does any one know even in terms of such a variable unit as a per cent just what is necessary in order to be able to continue with profit the work to higher levels? If 60% were made the passing standard, the 75%-minded teachers would immediately begin to give harder exams and quizzes; if 75% were made the passing standard, the 60%-minded teachers would begin to give easier exams and quizzes. In any event either group would go on passing about the *same per cent* of their classes as formerly. Now such a subjective system of units is quite unscientific; it is indefensible on any ground.

These "constant errors" or "systematic biases" on the part of teachers become reflected in school marks. The variability of school marks has been the subject of extensive investigation by Johnson,[1] Kelly,[2] Miles,[3] Starch and Elliott,[4] Dearborn,[5] and many others. The following table taken from Johnson is typical.

[1] JOHNSON, F. W. — "A Study of High School Grades," *School Review*, 19 : 13–24, Jan., 1910.

[2] KELLY, F. J. — *Teacher's Marks, Their Variability and Standardization.* TEACHERS COLLEGE CONTRIBUTION TO EDUCATION, VOL. 66, 1914.

[3] MILES, W. R. — *Comparison of Elementary and High School Grades,* UNIVERSITY OF IOWA STUDIES IN EDUCATION, VOL. 1, no. 1, 1910.

[4] STARCH, D. and ELLIOTT, E. C. — "Reliability of Grading High School Work in English," *School Review*, 20 : 442–457, Sept., 1912. — "Reliability of Grading High School Work in Mathematics," *School Review*, 21 : 254–259, April, 1913. — "Reliability of Grading High School Work in History," *School Review*, 21 : 676–681, Dec., 1913.

[5] DEARBORN, W. F. — "School and University Grades," *University of Wisconsin Bulletin*, no. 368, 1910.

TABLE 4

DISTRIBUTION OF MARKS IN THE UNIVERSITY OF CHICAGO HIGH SCHOOL FOR 1908–1909 (From Johnson)

Department	Number of Marks	F		D		C		B		A	
		Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
Latin and Greek	886	94	10.6	143	16.1	282	31.8	208	23.5	159	17.9
German	416	35	8.4	81	19.5	110	26.4	119	28.6	71	17.1
French	475	52	10.9	89	18.7	157	33.0	133	28.0	44	9.3
English	1514	235	15.5	329	21.7	497	32.8	354	23.4	99	6.5
History	825	67	8.1	131	15.9	258	31.2	248	30.0	121	14.7
Mathematics	1466	212	14.5	370	25.2	405	27.6	310	21.1	169	11.5
Science	672	56	8.3	113	16.8	186	27.7	219	32.6	98	14.6
Shop and Drawing	867	76	8.8	122	14.1	289	33.3	287	33.1	93	10.7
Domestic Science	176	10	5.7	4	2.3	48	27.3	91	51.7	23	13.1
TOTAL	7297	837	11.5	1382	18.9	2232	30.6	1969	27.0	877	12.0

An inspection of this table shows wide variability between departments. A child taking Latin and Greek has almost three times the chance of getting *A* as in English. English teachers, on the other hand, fail nearly three times as many as domestic science teachers. Bunching together the marks of a department causes differences between teachers to disappear. Two teachers in the same department showed the following percentages of marks in Johnson's study.

	F	D	C	B	A
First teacher	7.5%	15.7%	47.3%	22.3%	7.2%
Second teacher	4.5%	5.7%	23.8%	30.3%	35.7%

Professor Johnson assures me that the school did not have ability grouping at the time when this tabulation was made. Since the two classes were sectioned without regard to ability, the chances are that they were very nearly equal in ability and much nearer

in achievement than the above figures would lead one to believe.

To be really conclusive it would be desirable to compare two sets of final marks given to a class at the same time by different teachers, just as two sets of examinations have been corrected by different people. This, of course, is not possible, as a class is not taught by two teachers in the same subject at the same time. F. J. Kelly was able, however, to compare the final grades of a class in two successive semesters taught by different teachers in Hackensack, New Jersey. Concerning this experiment he says:[1]

> There are four ward schools which send their pupils at the completion of the sixth grade to a common seventh grade where there is departmental teaching. It seemed to me important to determine how far the pupils from each school maintained their relative positions in the common seventh-grade classes. If there should be found a difference in the amount of the increase or decrease of marks from sixth to seventh grade among the four school groups as wholes, it would be possible to measure with some degree of security the difference in standard between a given mark in one sixth grade and the same mark in another sixth grade. The marks recorded on the cards in Hackensack are letters as follows: "E," "G," "F," and "P," for excellent, good, fair, and poor, respectively.

Kelly reports such differences as these:

> For work which the teacher in School C would give a mark of "G" in language, penmanship, or history, the teacher in School D would give less than a mark of "F." A pupil whose monthly report card in School C had been a "G" card, on the whole, would be dismayed by receiving an "F," when he moved to School D.

Since the causes of such inadequacies in term marks or examination are most significantly described by means of technical terms, I shall postpone their discussion until certain technical matters in statistics have been given exposition. The foregoing evidence has been adduced to bring into bold relief the defectiveness of the measurements that are now being made and used in high schools. It is true that oftentimes the mere presence of a defect

[1] *Op. cit.*, p. 7.

is not a sufficient argument for remedying the defect. Some take the attitude, "What difference does it make if we do not try to improve our measures of pupil-achievement in school and continue to use the measures that we were accustomed to using in the past?" One reason why our present system of measures is so pernicious is that it does injustice to individuals. Wood[1] has demonstrated that the College Entrance Examination Board examinations deny entrance to 1,279 out of 10,000 applicants for college whom a similar equivalent examination would admit to college. This is an instance where gross injustice in the life careers of numbers of young men and women occurs yearly. Similar injustice also occurs daily in school. Using Wood's same figures, on the basis of traditional examinations or teachers' offhand judgments about 10 out of 100 are placed in the *B* section in a class in school who belong in the *A* section, and vice versa. This separation of the brighter and duller which occurs every time there is promotion or failure or segregation in school is a matter affecting the destiny of every one in school and should be done as accurately as possible. Seen in this light it is important that expense be not spared in making sure that individuals are treated rightly in the critical points in their school careers and that vocational advice shall be based on the best evidence obtainable and not on subjective opinion or judgment.

But the big injustice of imperfect measurement lies closer to the learning process than this. The evidence which a pupil receives of his own past learning is a potent factor in his present learning. Kirby,[2] Chapman and Feder,[3] Hurlock[4] and others have demon-

[1] Cf. *Measurement in Higher Education*, p. 124 f.

[2] KIRBY, T. J. — "Practice in the Case of School Children," *Teachers College Contributions to Education*, no. 58, 1913.

[3] CHAPMAN, J. C. and FEDER, R. B. — "The Effect of External Incentives on Improvement," *Journal of Educational Psychology*, 8 : 469–474, Oct., 1917.

[4] HURLOCK, E. B. — *The Value of Praise and Reproof as Incentives for Children*, ARCHIVES OF PSYCHOLOGY, NO. 71, July, 1924.

strated experimentally the value of praise as an incentive for school work. Miss Hurlock states as the conclusion of her experimentation concerning the value of praise, censure, and ignoring as incentives in school work:

> The results as a whole point conclusively to the fact that regardless of the factors of age, sex, initial ability, or accuracy, praise is decidedly the most effective of the three incentives here investigated. Reproof, when first used, seemed to be about equal in value to praise, but with continued use its effectiveness showed a decided decline. To ignore children in a group where the other members are receiving some incentive seems to be psychologically bad. The longer they are ignored, the less improvement do they show, even in spite of the opportunity to improve which comes from continued practice in one form of work. Hence, we may conclude that praise is the best form of incentive to use, no matter from what angle we may regard the matter.

School marks constitute a very important kind of praise or blame. A pupil who receives an *A* or a *B* in a subject is inclined to consider himself a good student, whereas one who receives a *D* or *E* considers himself a poor student or a failure. This expansion or contraction of the personality in the direction of study has its consequent effect upon further work. It is imperative therefore, that we give pupils as correct a judgment as possible of their achievement in order that those whose ability warrants may be encouraged to pursue their course with zeal and ardor, while those who are not likely to profit by their efforts may be justly discouraged from further endeavor, and perhaps be led to turn their talents in more fruitful directions.

I have criticised school examinations and marks as they are. From a wider point of view the very way in which school marks are expressed is not the best for the purpose. Letter marks such as *A*, *B*, *C*, *D*, and *F* are *relative;* what is desired is some system of *absolute* measurement. Suppose Charles takes home a report of *B* in English. What do his parents know about his achievement? They know nothing except that the teacher thought Charles

could have done slightly better than he did do, but that he is considerably above passing. They know that he was probably surpassed by a few gifted members of his class, but that, on the other hand, he surpassed in achievement the majority of the class. They do not know how his achievement compares with last year's class; with that of a class in a neighboring high school in some other city; or even know how it compares with his own achievement at the time of the last report. Presumably he has learned something in English in the meantime, since he has been to school regularly and *holds the same relative position* in his class, as evidenced by the mark *B* received in his previous report. But one should not have to deduce learning from such circumstantial evidence. Mere attendance at school is not a guarantee of learning. English classes have been known to spend a term with no apparent improvement in spelling, ability to punctuate, to capitalize, to use correct English, to write acceptable letters, or to improve in any of the measurable evidences of composition ability. Charles may be in a class which, under poor teaching, has grown in carelessness in the use of the English language, but the mark of *B* would give no evidence of the fact.

The percentage system by which traditional examinations are corrected is also a relative scale. It is a scale which is made to fit every examination given, whether hard or easy. Eighty per cent tells a parent practically nothing. With a hard examination 80% means a high grade — with an easy examination 80% means a low grade. Until one has seen the examination it is not known what achievement 80% represents.

An absolute mark is represented by the point on a scale which extends from the level where the first skill is obtained or the first bit of information is learned in a subject until a level is reached which is as difficult as anything offered in that subject. This scale should extend by steps which are practically equal in units of difficulty and by means of which the level of difficulty or skill or

achievement reached by a pupil may be stated in absolute terms — that is, units of the scale. Such a test in algebra would begin with an example of extreme ease and simplicity and would progress with examples of increasing difficulty until the most difficult type of exercise taught would be included. Pupil X taking the test would be able to do the easy ones, but somewhere would find examples of a difficulty that is beyond his ability. This place in the test where he begins to fail would define his achievement in the subject. This achievement may be reported in terms of number of examples which pupil X can successfully do on the test, or better, in terms of the length of time spent in study by the average boy who takes algebra before he has an achievement equal to the achievement of pupil X. That is, a pupil X may have a score equivalent to that made by the average pupil who has studied algebra four months. In this case 4 would be a very convenient and illuminating mark to use as it signifies a definite and absolute amount. If, now, pupil X has studied algebra only three months, he at once knows that his achievement is better than that of the average pupil studying algebra. By means of such a set of marks, definite goals and standards may be established for promotion, for college entrance, for special honors, and the like. Such a test is known as a standardized test and it is the purpose of this book to describe such standardized tests as are extant and to give directions for their use and interpretation. Such a mark as above described is known as a standardized mark. Its use is described on p. 500. It is only by means of an absolute measure of achievement that we can get away from the morass of individual injustice that is practiced every day in our high schools. A great deal of research must be done before we have adequate instruments for assigning absolute marks in all subjects as above described. The need is apparent. Much has been accomplished toward making measurement in secondary education more exact, much still lies ahead.

REFERENCES

BLISS, D. C. — "High School Failures," *Educational Administration and Supervision*, 3 : 125–138, March, 1917.

CARTER, R. E. — "Correlation of Elementary Schools and High Schools," *Elementary School Teacher*, 12 : 109–118, Nov., 1911.

CHAPMAN, J. C. and FEDER, R. B. — "The Effect of External Incentives on Improvement," *Journal of Educational Psychology*, 8 : 469–474, Oct., 1917.

CHAPMAN, J. C. and RUSH, G. P. — *The Scientific Measurement of Classroom Products*, ch. 1. Silver, Burdett & Company, Boston; 191 pp., 1917.

COMIN, R. — "Teachers' Estimate of the Ability of Pupils," *School and Society*, 3 : 67, Jan. 8, 1916.

DEARBORN, W. F. — *Relative Standing of Pupils in the High School and in the University*, UNIVERSITY OF WISCONSIN BULLETIN NO. 312, 1909.

— *School and University Grades*, UNIVERSITY OF WISCONSIN BULLETIN, NO. 368, 1910.

FOSTER, W. T. — "Scientific versus Personal Distribution of College Credits," *Popular Science Monthly*, 78 : 388–408, April, 1911.

GRAY, C. T. — *Variations in the Grades of High School Pupils*, Warwick and York, Baltimore, 1913.

GREGORY, C. A. — *The Fundamentals of Educational Measurement*, ch. 5. D. Appleton & Company, New York; 382 pp., 1922.

HURLOCK, E. B. — "An Evaluation of Certain Incentives Used in School Work," *Journal of Educational Psychology*, 16 : 145–159, March, 1925.

— *The Value of Praise and Reproof as Incentives for Children*, ARCHIVES OF PSYCHOLOGY, NO. 71, New York, July, 1924.

JOHNSON, F. W. — *Administration and Supervision of the High School*. Ginn & Company, Boston, 1925.

— "A Comparative Study of Grades of Pupils from Different Elementary Schools in Subjects of the First Year, High School," *Elementary School Teacher*, 11 : 63–78, Oct., 1910.

— "A Study of High School Grades," *School Review*, 19 : 13–24, Jan., 1910.

KELLY, F. J. — *Teachers' Marks, Their Variability and Standardization*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, VOL. 66, Bureau of Publications, Teachers College, Columbia University, New York, 83 pp., 1914.

KIRBY, T. J. — *Practice in the Case of School Children*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 58, Bureau of Publications, Teachers College, Columbia University, New York, 1913.

MEYER, M. — "The Grading of Students," *Science*, 28 : 243–252, April 21, 1908.

MILES, W. R. — *Comparison of Elementary and High School Grades*, UNIVERSITY OF IOWA STUDIES IN EDUCATION, VOL. I, NO. 1, 1910.

MONROE, W. S. — *Measuring the Results of Teaching*, ch. 1. Houghton Mifflin Company, Boston; 297 pp., 1918.

— *Written Examinations and Their Improvement*, ch. 1. UNIVERSITY OF ILLINOIS BULLETIN, NO. 9, VOL. 20; 71 pp., 1923.

MONROE, W. S., DEVOSS, S. C., and KELLY, F. J. — *Educational Tests and Measurements*, ch. 1. Houghton Mifflin Company, Boston; sec. ed., 521 pp., 1924.

MONROE, W. S. and SOUDERS, L. B. — *The Present Status of Written Examinations and Suggestions for Their Improvement*, ch. 2. UNIVERSITY OF ILLINOIS, BULLETIN NO. 17, VOL. 21; 77 pp., 1923.

RUCH, G. M. — *Improvement of the Written Examination*. Scott, Foresman & Company, Chicago, 1924.

STARCH, D. — *Educational Measurements*, ch. 2. The Macmillan Company, New York; 202 pp., 1918.

— *Educational Psychology*, ch. 22. The Macmillan Company, New York.

STARCH, D. and ELLIOTT, E. C. — "Reliability of Grading High School Work in English," *School Review*, 20 : 442–457, Sept., 1912.

— "Reliability of Grading High School Work in History," *School Review*, 21 : 676–681, Dec., 1913.

— "Reliability of Grading High School Work in Mathematics," *School Review*, 21 : 254–259, April, 1913.

WOOD, B. D. — *Measurement in Higher Education*. World Book Company, Yonkers, 1923.

— "Measurement of College Work," *Educational Administration and Supervision*, 7 : 301–304, Sept., 1921.

CHAPTER III

METHODS OF IMPROVING EXAMINATIONS

Types of Examinations

Perhaps the best introduction to the problem of measuring in secondary education is to review the methods that have been developed for improving ordinary school examinations. The devices and procedures which will be described give suggestions which most teachers are glad to try even though they are not immediately willing to discard their traditional methods of examining. Noting these devices will serve as a good introduction to our study of certain standardized tests, for most standardized tests have adopted these newer forms of test construction. One way of improving tests is to construct a test which permits only one correct answer so that there can be no disagreement in the scoring. Spelling tests have always been objective in that there can be no disagreement in the correctness of the spelling of any word. The answer to a problem in algebra is either right or wrong. But history and civics tests have not been so constructed as to admit of such objectivity of scoring. It is the purpose of this chapter to give suggestion as to the construction of objective tests. Seven kinds of tests will be described.

I. The True-False Examination

In making an examination the usual way is to ask a question to which the pupil gives an answer, *e.g.*, why does a laundry iron use more electricity than a 40-watt lamp? In the true-false test one makes such a statement as: A laundry iron uses more electricity than a 40-watt lamp because it offers less resistance to

the flow of current. The pupil has to indicate whether this statement is true or false. In constructing a true-false test the following rules should be observed:

1. Prepare a list of statements which cover in some detail the portion of the subject on which the pupils are to be examined. If a large amount of ground is to be tested, take care that the questions are a representative sampling of the material. After such a list has been prepared, some of the statements can easily be changed so that they are false. The untruth of a statement should not be too obvious or it will be worthless for testing. For instance, if the above statement read "A laundry iron uses more electricity than a 40-watt lamp because it is heavier," it would be obviously so ridiculous as to make it a poor test item. Also statements should be selected which require an acquaintance with the subject in order to determine their truth or falsity.

2. Some of the pitfalls to be avoided in making a true-false statement are:

a. Avoid long statements. Weidemann[1] suggests that good statements can be constructed with from 10 to 20 words. Longer statements are sometimes necessary.
b. Avoid ambiguous statements.
c. Avoid trivial statements.
d. Avoid suggestive statements.
e. Put statements positively where possible.
f. See that one statement does not answer a preceding one.
g. Avoid dependent clauses (especially those which determine the truth or falsity of the statement).
h. Avoid compound sentences containing two separate ideas either of which may be false. If either half of a compound sentence is false, the whole item must be considered false.[2]

[1] WEIDEMANN, C. C. — *How to Construct the True-False Examination*, p. 78, 1926.

[2] Both BRINKLEY, *Values of New Type Examinations in the High School*, and WEIDEMANN (*op. cit.* pp. 43–59) give a long list of rules, with illustrations, for avoiding pitfalls in constructing statements in a true-false test. These authorities should be consulted.

In general the rules governing English composition apply to the construction of true-false statements.

3. Weidemann[1] points out that there are "specific determiners" in the true-false test which the clever pupil may use to guide him in determining his response. Such are these:

2 out of 3 "always" or "never" statements are *false*.

2 out of 3 degree or comparison statements are *true*.

2 out of 3 cause or reason clauses are *false*.

For instance, *All sheep have wool,* is a statement whose specific determiner is *all*. Two out of three such statements are false in the long run.

The teacher who is making a true-false test should not avoid these specific determiners, but should make sure that she is not consistently using them so that the response is in one direction.

4. In a true-false examination some of the statements should be true and some false. If all of the statements were false, the crafty pupil would soon catch on and it would be no test for him. The number of true statements should be approximately the same as the number of false statements. They should be so arranged that there is a chance order of true statements and false statements.

5. The number of statements depends in part on the subject matter to be tested and on how the examination is used. If the test covers only a short section or chapter, one naturally could not construct an unlimited number of valid items. Likewise if the true-false test is part of a more comprehensive test containing other kinds of items, fewer true-false statements are required than if the true-false test stands alone. In general true-false tests are not very reliable unless one hundred or more statements are included. But one should not stretch limited subject matter to a test of this length, neither should one feel obliged to put one hundred true-false statements into an examination containing other

[1] *Op. cit.*, pp. 68 ff.

types of questions. On the other hand, one should never give a true-false test of ten or twenty items alone and consider the results of value in determining the achievement of individual pupils.

6. The true-false test should be mimeographed or printed so that each pupil will have a copy. Weidemann[1] has determined the best way of printing the true-false test, the skeleton form of which is indicated in the following scheme.

......1. ....

......2.

......3. ....

......4. ....

......5. ....

etc.

On the dotted line before the number of each statement the answer is to be indicated by the pupil. If the pupil believes the statements to be true, he places a "+" on the dotted line before the statement. If he believes the statement to be false, he places a "o" on the dotted line before the statement.

If it is not possible to give the pupils mimeographed copies of the tests, it is permissible to read the statements to them provided their understanding of spoken English is good. If the true-false statement is given orally, it should be done very deliberately. Each statement should be read twice and plenty of time allowed between statements to give the class opportunity to think out their answers. If the test is given orally, give each child a blank sheet

[1] *Op. cit.*, pp. 24 f., 102.

of paper on which to write in a column consecutive numbers corresponding to the numbers of the statements of the test. Opposite each number he will write a plus or a zero sign to indicate his judgment as to the truth or falsity of the statement. The disadvantage of the oral plan is that the pupils do not have a satisfactory opportunity to study the statements. Also the class may give some indication of the answer if a statement appeals to some pupils as being ridiculous.

7. Weidemann[1] suggests that the following directions be given with a true-false test.

Directions: About one half of the following statements are true and about one half are false. Mark each true statement with a plus sign (+) on the dotted line at the left of the statement. Mark each statement that is partly or wholly false with a zero (o) on the dotted line at the left of the statement. Do not mark statements which you do not know.

Mark the statements in order.

DO NOT GUESS. Guessing reduces your score.

Ask no questions.

For some time the question as to whether pupils should be instructed to guess or not to guess was debated. Recently Ruch[2] has produced evidence that the most valid and reliable results are obtained on true-false tests by instructing the pupils not to guess.

8. In order to do away with the influence of guessing the true-false test should be scored right minus wrong (R − W). It must be evident that a pupil can get a considerable score on a true-false test when he is totally ignorant of the subject matter. For instance, a pupil might go through the test and mark every statement true, in which case he would get about half of them right, whereas of course his score should be zero. By using the formula

[1] *Op. cit.*, p. 98.

[2] Ruch, G. M., and DeGraff, M. H. — "Corrections for Chance and 'Guess' *vs.* 'Do not Guess' Instructions in Multiple-Response Tests," *Journal of Educational Psychology*, 17: 368–375, Sept., 1926.

R — W in such a case the score would be zero or very near zero.[1] To give a concrete example, suppose one has a test of sixty statements. Pupil A taking the test omits thirteen statements and gets seven statements wrong. Weidemann[2] suggests the following form:

Number of statements	60	
Number of omissions	13	
Rights and Wrongs	47	by subtraction
Number of Wrongs	7	
Number of Rights	40	by subtraction
Number of Wrongs	7	
Score is (R — W)	33	by subtraction

An error in scoring when using the R — W formula is to take the number right the same as the number tried. One should remember that the number right *plus* the number wrong equals the number tried.[3]

[1] There is never a sharp break between a statement which is known and a statement which is not known. Rather the known shades off into the unknown and there is an area in which a pupil dimly knows whether the statement is true or false. These statements in this twilight zone will be marked correctly more often than not and hence a score will represent both those statements whose truth is fully known and some statements whose truth is only dimly known. Probably some individuals are better able than others to draw on the resources of their scraps and shreds of information and to make lucky guesses. However, these lucky guesses which represent a skilful use of stray associations are certainly worth the credit which they gain.

[2] *Op. cit.*, p. 98.

[3] A controversy has raged over the correct formula to be used in scoring true-false tests. It is now generally conceded, following the experimental work of Wood and Ruch, that the R — W formula is to be preferred. Holzinger * has shown that it is sufficient merely to count the number of correct responses in scoring a true-false test, providing all of the elements have been answered. In this case the scores will be identical with R — W scores *insofar as they place the members of the class in the order of merit.* Of course it will be realized that scores determined by the R — W method and scores determined by the number right are not interchangeable, the latter set being consistently higher than the first set.

* Holzinger, K. J. — "On Scoring Multiple-Response Tests," *Journal of Educational Research*, 15: 445–447, Oct., 1924.

9. In scoring a set of papers one should make a scoring key. This may be a strip of manila paper with the correct responses (+ or o) indicated, care being taken to space them exactly even with the dotted lines on which the pluses and zeros are written on the test papers. The scorer brings this key as closely as possible to the responses and on the same level so that there will be little eye strain. It has been found best to check *wrong* statements with a red or blue pencil (the check may be only a straight line about half an inch long) and to indicate *omissions* with a red or blue zero. All *right* answers are left unmarked.

A variation of the true-false test is called the "yes-no" test. Instead of giving a statement a question is asked which may be answered yes or no. An example is: "Do you measure the amount of electricity by a voltmeter?" Probably the true-false test and the yes-no test are equally satisfactory. The rules for giving and scoring the true-false test may be followed in using the yes-no test. Following are examples of yes-no questions from two standardized tests:

HAGGERTY READING SIGMA 3: FORM A, TEST 2[1]

Directions: In the following pages are some sentences. Each sentence asks a question which can be answered by *yes* or *no*. Mark the right answer to these questions by drawing a line under the *yes* or the *no*. Do not mark both *yes* and *no*. Mark only the right answer.

1.	Is snow white?	YES	NO
2.	Are elephants plants?	YES	NO
3.	Can a pupil respond to a question?	YES	NO
4.	Are multitudinous defects desirable?	YES	NO
5.	May a hamlet be located in a province?	YES	NO
6.	Do caravans always move with great speed?	YES	NO
7.	Is day always preceded by night?	YES	NO
8.	Can a boy be absorbed in a performance?	YES	NO
9.	Do vicious men plan revenge?	YES	NO
10.	Are all experiences humiliating?	YES	NO

[1] Copyright 1920 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

ON TRAINING C

OTIS GROUP INTELLIGENCE SCALE — ADVANCED EXAMINATION [1]
TEST 10. MEMORY

(In this test a story is first read aloud to the class and then the pupils taking the test answer the following questions by underlining YES, NO, or DIDN'T SAY.)

1. Was the king fond of hearing stories?	YES	NO	DIDN'T SAY
10. Did the king then order another man to tell him a story?	YES	NO	DIDN'T SAY
20. Was a very small hole left between the bricks near the ground?	YES	NO	DIDN'T SAY

At the end of this chapter is a specimen true-false test. The pluses and zeros are those which a pupil might have given. The wrong ones have been checked. The *Right-minus-Wrong* method is used in scoring. The score on the test is 26, which was found as follows:

The number of correct statements is	34
The number of wrong statements is	8
Subtracting *Wrongs* from *Rights* we obtain a score of	26

Sometimes there are more wrongs than rights, in which case the score would come out less than zero. But it is customary to give no score less than zero — one should not be penalized for bad luck in guessing.

II. The Multiple-Choice Examination

An illustration of a unit in the multiple-choice test is the following:

A widely known poem written by Goldsmith is (1) Marmion, (2) The Deserted Village, (3) The Ancient Mariner, (4) Childe Harold. ( )

In this test the pupil is given several choices, one of which makes the statement a true statement. His task is to place in the parenthesis the number of the choice which makes the statement true. The marking of this test is objective since only one of the answers is correct.

[1] Copyright 1918 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

In making this type of examination none of the choices should be too obviously incorrect; for example: The man who discovered America was: Henry Ford, Christopher Columbus, Thomas Edison, Babe Ruth. Here the three choices that are not correct are too ridiculous to make the statement valuable in a test. The correct answer should not always be found in the same position. Sometimes it should be first, sometimes last, and sometimes in an intermediate position. In general, four or more choices should always be given. If fewer than four choices are allowed, chance plays a significant part in determining the score. A multiple-choice test, like the true-false test, should consist of at least a hundred statements. A multiple-choice test cannot be given orally. It must be printed or mimeographed. The most economical way of arranging the test is to number each of the alternatives in a statement and let the pupil indicate the alternative he chooses as the correct one by writing its number in a parenthesis provided at the right of the statement. This makes the scoring of the test easier than when the correct alternative is merely underlined. The score on the multiple-choice test is the number of correctly answered items. Since there are always four or more choices, chance plays a very little part in determining the score. At the end of this chapter are specimen multiple-choice tests. The statements have been answered as though a pupil had taken the test. Wrong responses have been checked. The score on the "Plot" test is 33, and on the "Characters" test is 13.

Both the true-false and the multiple-choice tests may be used in all high school subjects. Examples of multiple-choice statements taken from various standardized tests are presented:

RUCH-POPENOE GENERAL SCIENCE TEST [1]

1. Pneumonia is a disease of the heart liver lungs brain stomach muscles kidneys

[1] Copyright 1923 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

10. When a liquid contains all the dissolved substance possible the condition is termed osmosis permeability fusion neutralization reduction transpiration saturation
20. The separation of liquids and solids by evaporation and condensation is called solution distillation diffusion fusion transpiration reduction pasteurization

RUCH-COSSMAN BIOLOGY TEST[1]

1. Organisms that secure their food supply at the expense of other living organisms are called saprophytes angiosperms antheridia spirogyra gymnosperms parasites bryophytes
10. The stage in the development of an insect which immediately follows fertilization is the adult pupa cocoon chrysalis gastrula egg blastula
20. The intestines are held in place by ligaments tendons mesenteries convolutions pseudopodia vertebrae muscles

POWERS GENERAL CHEMISTRY TEST[1]

1. Substances which hasten a chemical action without themselves undergoing chemical changes are called catalysts electrolytes ionogens allotrops colloids
10. Cotton fibre is nearly pure levulose dextrose lactose cellulose invertose
20. The oxygen of commerce is made from water potassium chlorate barium peroxide liquid air sodium peroxide

WHITE LATIN TEST[1]

Part I Vocabulary. *Directions:* Place in parenthesis on right margin the number of word which you think correct.

1. vadeo	(1) conquer (2) see (3) seem (4) come	( )
10. ita	(1) so (2) go (3) likewise (4) way	( )
20. gero	(1) tribe (2) almost (3) wage (4) give	( )

Part II Sentences. *Directions:* Put a cross before the letter of the sentence you think is correct.

1. Homo est magnus.
   (*a*) The home is large. (*c*) The man is a magnate.
   (*b*) There is a large home. (*d*) He is a large man.

[1] Copyright 1924 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

10. Prae eratne aliquis equitatui quem auxilia Caesari socii miserant?
    (*a*) Was anybody in charge of the cavalry which the allies sent to aid Caesar?
    (*b*) Who, pray, was in charge of the horsemen whom the allies sent to the aid of Caesar?
    (*c*) Was any one in command of the cavalry which the allies had sent to aid Caesar?

20. Sua interesse nulli hosti parcere videt cum se caedis infinitae paeniteat.
    (*a*) He sees that to spare the enemy would be of no interest to him, although he would regret an endless slaughter.
    (*b*) Although he would regret an endless slaughter, he sees that it would be to his interest to spare no enemy.
    (*c*) He sees that no enemy would be interested in sparing him, although he might be sorry for an endless slaughter.

HOME ECONOMICS INFORMATION TESTS [1]

If the bedroom is in the north side of the house, a good color to use on the walls would be
(1) red
(2) blue
(3) a light tan

Cotton is
(1) an animal fibre
(2) a vegetable fibre
(3) a mineral fibre

Bread flour is obtained from
(1) wheat
(2) corn
(3) barley

PRESSEY TEST IN LATIN SYNTAX [2]

1. They throw spears.
   Hastīs jaciunt. Hastam jaciunt.
   Hastae jacuint. Hastās jaciunt.

11. He was sent by Caesar.
    A Caesare missus est. Caesare missus est.
    A Caesarem missus est. A Caesarō missus est.

[1] Published by Teachers College Bureau of Publication.
[2] Published by Public School Publishing Company.

21. Part of the soldiers were sent.
Pars mīlitēs miss est. Pars mīlitum missa est.
Pars dē mīlitibus missa est. Pars mīlitibus missa est.

III. The Matching Test

At the end of this chapter on p. 48 is a specimen matching test. One takes this test by writing at the right of the lettered item the number of the item which the term defines. The directions indicate how to take the test. The first four items have been marked to show the method. The score is the number of correct matchings. This type of test may be made for all high school subjects by the skilful teacher. Although it has not been experimentally determined, I should say that not more than fifteen or twenty items should be included in each column, as a test with more than that number becomes much more difficult mechanically. Below are samples.

BARR DIAGNOSTIC TEST IN AMERICAN HISTORY[1]

TEST V

2. In the left-hand column below is a list of causes. In the right-hand column is a list of results. Take your pencil and draw a line connecting each cause with the proper result.

(*a*) The sinking of the Maine	(*a*) Territorial claims on Oregon
(*b*) The growth of industrial combinations	(*b*) The Mexican War
(*c*) The annexation of Texas	(*c*) The growth of abolition sentiment
(*d*) The Dred Scott decision	(*d*) The Spanish-American War
(*e*) The Lewis and Clark expedition	(*e*) The Sherman Anti-Trust Law

STEVENSON LATIN VOCABULARY TEST[1]

Directions: Notice the word *certes* at the top of column II in the test below. In column I are some English words arranged alphabetically. Find among these words the one which gives the meaning of *certes*, and write its number in the parenthesis in front of *certes*.

[1] Published by Public School Publishing Company.

COLUMN I		COLUMN II
1	book	( ) certes
2	certain	
3	curtain	( ) similis
4	daughter	
5	day	( ) deus
6	fear	
7	God	( ) liber
8	liberty	
9	seven	( ) filia
10	silver	
11	similar	( ) septem
12	summer	

IV. The Completion Test

The completion test consists of a number of statements from which words have been omitted. Pupils taking the test must supply the missing words. Care must be taken not to omit too many words as the test becomes difficult when the omissions are numerous. The best plan is to prepare a list of important statements covering the portion of the subject on which the pupils are to be tested and then from these statements to strike out a significant word or words. In every case the words omitted should be the only ones that will fit in the blank spaces. The pupil's score is the number of blank spaces filled in correctly. Oftentimes, even though great care is taken, a word other than the one stricken out will fit in the blank space. In this instance, the scorer will have to use his judgment as to whether it should receive credit. In giving a completion test each pupil must be provided with a mimeographed or printed copy of the examination.

Below is a sample completion test.

RUCH-COSSMAN BIOLOGY TEST [1]

1. To test for grape sugar, crush or scrape some of the material and boil in water. While the preparation is still boiling, add a few drops of ____. A ____ precipitate indicates the presence of grape sugar.

[1] Copyright 1924 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

3. Flowers are adapted in three ways to avoid self-fertilization; first, stamens and the ____ may be found in ____ flowers; secondly, the ____ may ripen before the ____ is ready to receive it; and thirdly, the ____ and ____ may be so situated that ____ cannot easily take place.

There are other types of objective tests less well known and less often used. They are also more difficult to construct than the ones already described. However, they may be used in many high school subjects.

V. The Analogies or Mixed Relations Test

Examples of the analogies test in geometry are:

Point : line = line : ?
Area : circumference = πr^2 : ?
Triangle : 180° = hexagon : ?

or in chemistry:

1 : hydrogen = ? : nitrogen
sodium sulphate : Na_2SO_4 = hydrogen sulphide : ?

VI. The Similarities Test

This test may be used wherever one wants to test knowledge of classification.

Example:

OTIS GROUP INTELLIGENCE SCALE [1]

TEST NO. 8

Directions: Find the way in which the first three things on a line are alike. Then look at the five other things on the same line and draw a line under the one that is most like the first three.

1. red, white, green	rose, paper, grass, soft, blue
2. ivory, snow, milk	butter, rain, cold, cotton, water
3. joy, anger, fear	habit, memory, hate, life, hearing

VII. The Classification Test

A somewhat similar scheme which Terman uses and which Pressey has more strikingly called a "Cross-Out Test" is as follows.

[1] Copyright 1918 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

TERMAN GROUP TEST OF MENTAL ABILITY[1]

TEST NO. 9

Directions: In each line cross out the word that does not belong there. Cross out just one word in each line.

1. Frank James John Sarah William
9. mechanic doctor lawyer preacher teacher
18. give lend lose keep waste

Advantages of the New Types of Examinations

1. The new type examinations are more objective in their scoring, since the responses of the pupils are controlled, and since there can be no doubt as to the correctness or incorrectness of their responses.

2. They may be very comprehensive and can be made to cover a great deal more material than the old type of examination. Since the pupil does very little writing, he can devote his time to thought and thereby can answer a great many more questions than when he has to write his answers out at length.

3. They are very easy to score as compared with the older type of examination. Notwithstanding that the tests are more comprehensive, they may be scored in less time and with less labor. This is of itself a sufficient reason to make them attractive to all teachers.

4. Pupils like these new types of tests. There is no question as to the accuracy of the marks they receive. There is no chance for the teacher to show favoritism or personal bias. The pupils, being relieved of much writing, find these tests less tiring.

5. They are more educative for the pupils. Let your pupils score their own tests directly after they take them by exchanging papers with their neighbors. It is then that their interest is at white heat. After the papers have been scored the questions may be discussed. The pupils readily see where they need further study.

[1] Copyright 1920 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

Limitations of the New-Type Examinations

1. The pupil does not have an opportunity to show his ability to organize his thought. He has nothing to do except check over the truth or falsity of the statements or fill in the missing words of incomplete sentences. Pupils miss the valuable experience of making comparisons, giving explanations, or giving definitions. They are not asked to summarize the material or to make applications of principles, and of course these are valuable abilities which we do not want to neglect.

2. These new types of examinations are not diagnostic in that they do not tell where the pupil's reasoning process goes wrong or where he stops reasoning altogether and starts guessing.

The traditional examination therefore has its merits and these are sufficient to guarantee its continued existence. However, these newer types of examinations have much in their favor and if the teacher's purpose is simply to test the pupil she can well make use of them. Progressive teachers everywhere are discovering that true-false exercises are not only useful for testing, but are an effective teaching device.

OBJECTIVE TEST ON
DICKENS' *A TALE OF TWO CITIES*

BY

GEORGE C. WELLS, HELEN E. HAWKINS, AND KATHERINE E. JESSUP

School .... Date ....

Name .... Date of birth .... Age ......

The Author and His Works

Directions. Make a plus sign (+) on the line at the left of each statement you consider true. Make a zero (o) on the line at the left of each statement you consider false.

Samples. o 1. Shakespeare was an American author.

+ 2. Dickens wrote *A Tale of Two Cities*.

o 1. Charles Dickens was the son of an English nobleman.
\+ 2. Washington Irving was a contemporary of Dickens.
\+ 3. A passion for reading was awakened in Dickens by his mother.
o 4. During his boyhood, Dickens was a healthy lad.
\+ 5. Dickens did not find any interesting books in his father's library.
o 6. Dickens began the study of Latin at school.
o 7. Dickens' father was placed in prison for stealing.
\+ 8. At the age of twelve, Dickens supported himself.
√ o 9. Bob Fagin was met by Dickens in a blacking factory.
o 10. None of Dickens' relatives, except his father, ever lived in prison.
\+ 11. Dickens' father learned shorthand when he was about forty-five.
o 12. Dickens was admitted to the bar.
\+ 13. Dickens excelled his fellow workers in the use of shorthand.
√ o 14. Dickens was a reporter.
\+ 15. Catherine Hogarth married Dickens.
\+ 16. Dickens was the father of ten children.
\+ 17. America welcomed Dickens as one of the most popular writers of the day.

NOTE. — The checks indicate items which have been answered incorrectly. See p. 32.

o 18. Before returning home from America, Dickens became very popular by his discussion of international copyright.
\+ 19. Dickens' practice in amateur theatricals prepared him to be a dramatic reader.
√ + 20. Dickens seldom took long walks.
o 21. In dress, Dickens was noted for his quiet refined taste.
\+ 22. In England and in America, Dickens achieved fame as a reader.
√ + 23. Dickens was very poor financially during the last years of his life.
o 24. Before Dickens died, he expressed a wish to be buried in Westminster Abbey.
o 25. Dickens' novels are best known for the romance and history which they contain.
o 26. His first published book convinced the public that Dickens was a genius.
\+ 27. *Sketches by Boz* was the first book written by Dickens.
√ o 28. Dickens was famous as a writer at twenty-four.
\+ 29. *Pickwick Papers* was devoted to the life and manners of the middle and lower classes of society.

o 30. The death of Little Nell is described in *David Copperfield.*
+ 31. Two of the most noticeable characteristics of Dickens' works are humor and pathos.
o 32. Generally, in Dickens' novels, emphasis is laid on plot rather than on characters.
√ o 33. *American Notes* gave great offense in America.
o 34. Dickens wrote *The Birds' Christmas Carol.*
+ 35. Much of the material in *David Copperfield* is autobiographical.
+ 36. Paul Dombey is a famous character created by Dickens.
+ 37. Fagin is the name of a master thief in *Oliver Twist.*
o 38. Dickens created the character of Becky Sharp.
√ o 39. Dickens had much difficulty in deciding upon a title for the novel he called *A Tale of Two Cities.*
o 40. *A Tale of Two Cities* is characteristic of Dickens' style.
+ 41. *Nicholas Nickleby* was a protest against cheap boarding-schools.
√ + 42. *The Mystery of Edwin Drood* was never completed by Dickens because he lost interest in the subject.

Background and Setting

Directions: Write in each blank space the correct term.

1. The chief events in *A Tale of Two Cities* are based historically on the ____.
2. As the chief source of his historical material, Dickens used ____.
3. The principal scenes of *A Tale of Two Cities* are laid in London and ____.
4. At the time of the opening of the story, ____ was king of England.
5. The story begins about the time of the ____ war in America.
6. At the beginning of the story, ____ was king of France.
7. The "queen with a fair face" was ____.
8. In France the aristocracy was living at the expense of the ____.
9. ____ had complained of the people's wrongs in *The Deserted Village.*
10. Near London the roads were infested with ____.
11. Petty thefts and brutal murders were alike punishable by ____.
12. The atmosphere at the beginning indicates that the story will be ____.
13. The mail coach in which Mr. Lorry was traveling journeyed from London to ____.
14. The season of the year in which the story opened was ____.

15. On the day Mr. Lorry journeyed from London, the weather was ____.
16. One of the chief historical events described in *A Tale of Two Cities* is the storming of the ____.
17. Doctor Manette was imprisoned in the ____.
18. The last scene of the story is laid in ____.
19. Dickens obtained the main idea of this story from Wilkie Collins' drama, ____.

Plot

Directions: Write the number of the best ending for each sentence given below on the marginal line as shown in these samples:

1. *Macbeth* is a (1) short story (2) comedy (3) novel (4) tragedy (5) lyric poem 4
2. *Ivanhoe* was written by (1) Ruskin (2) Scott (3) Tennyson (4) Byron (5) George Eliot 2

---

1. The incident which opens the story takes place (1) on the road to Dover (2) in the wine shop (3) at the Gorgon's Head (4) at Temple Bar (5) in Old Bailey's 1 1
2. The horseman who came up to the coach in which Mr. Lorry rode was (1) a highwayman (2) an additional guard (3) a belated passenger (4) a messenger from Tellson's bank (5) a disguised spy 4
3. At Dover, Mr. Lorry was joined by (1) Jerry Cruncher (2) Sydney Carton (3) Lucie Manette (4) Dr. Manette (5) Mr. Barsad 3
4. The scene of the broken wine cask was introduced to (1) supply humor (2) reveal the condition of the people (3) show the carelessness of the market people (4) produce an atmosphere of mystery (5) show how the people were fed 2
5. When found by Mr. Lorry, Dr. Manette was (1) attending a patient (2) working with chemicals (3) making shoes (4) writing his memoirs of prison life (5) sleeping 3
6. Jerry Cruncher's fingers were always rusty from (1) picking locks (2) opening coffins (3) cleaning his boots (4) carrying packages (5) turning a grindstone 1✓
7. At Old Bailey's, Charles Darnay was being tried for (1) petty theft (2) treason (3) forgery (4) larceny (5) embezzlement 2

8. Darnay's acquittal was accomplished by (1) Sydney Carton (2) Mr. Lorry (3) Lucie Manette (4) Roger Cly (5) Jerry Cruncher 1

9. Darnay had been traveling between London and France because (1) he loved Lucie (2) he wanted to redress the wrongs perpetrated by his family (3) his uncle's life was in danger (4) he was afraid of the French peasants (5) he needed recreation 2

10. Mr. Stryver's rise in the legal profession was due to (1) his keen mind (2) his personality (3) Lucie Manette's influence (4) his wife's help (5) Sydney Carton's help 5

11. Dr. Manette moved to the corner near Soho Square to (1) have quietude (2) make shoes (3) make a scientific study of echoes (4) escape a prison sentence (5) entertain his friends 1

12. Miss Pross resented Darnay's visits to the Manette home because of (1) her jealousy (2) her affliction (3) his personality (4) Dr. Manette's weakness (5) Mr. Lorry's objection 1

13. Gaspard hated the Marquis because (1) he lived in a magnificent château (2) his carriage had killed Gaspard's child (3) he drank more than his share of the wine (4) he was cruel to his servants (5) he ordered Gaspard to be killed 4✓

14. The Marquis regarded Darnay with (1) affection (2) toleration (3) fear (4) gratitude (5) dislike 5

15. The Marquis was murdered by (1) Barsad (2) the mender of roads (3) Evrémonde (4) Gaspard (5) Defarge 4

16. Charles Darnay made his home in England in order to (1) earn his living (2) please his uncle (3) study English (4) help the poorer classes in England (5) spend his fortune 3✓

17. Mr. Stryver was prevented from offering himself in marriage to Lucie Manette by (1) Sydney Carton's disapproval (2) Mr. Lorry's advice (3) Dr. Manette's objection (4) Miss Pross' indignation (5) Darnay's anger 2

18. Sydney Carton told Lucie that he would (1) stop drinking (2) pay more attention to his work (3) give his life to keep one she loved (4) be neater in his habits of dress (5) never come to her home again 3

19. The purpose of Madam Defarge's knitting was to (1) provide clothes for the poor (2) earn money (3) give warning to spies (4) register names (5) unravel a ball of yarn 2✓

20. The name *Jacques* was given to those who (1) plotted secretly against the French nobility (2) sold wine (3) mended roads (4) collected taxes from the people (5) sat on the King's Bench 3✓

21. Dr. Manette wished Darnay to reveal his French name (1) before proposing to Lucie (2) on his wedding morning (3) under no conditions (4) in Mr. Lorry's presence (5) after talking with Stryver 2

22. The first important act of the revolutionists was the (1) arrest of Gabelle (2) punishment of the murderer of the Marquis (3) burning of the Marquis' château (4) capture of Foulon (5) storming of the Bastille 5

23. After the uprising in France, Mr. Lorry went to Paris to (1) rescue Darnay (2) take care of Tellson's business (3) see the country (4) visit his father (5) be with Miss Pross 2

24. Charles Darnay resolved to go to Paris to (1) accompany Mr. Lorry (2) help the common people (3) rescue a servant from prison (4) escape from arrest (5) protect his property 3

25. On his arrival in France, Darnay was (1) treated with consideration (2) beheaded (3) imprisoned (4) welcomed by Defarge (5) sent back to England 3

26. When Lucie and Dr. Manette arrived in Paris, they were provided for by (1) Solomon (2) Defarge (3) Mr. Lorry (4) the Vengeance (5) the farrier 3

27. Dr. Manette's first appeal to the tribunal was (1) successful (2) discredited (3) postponed (4) overlooked (5) censured 1

28. Miss Pross' long-lost brother Solomon proved to be (1) Mr. Lorry (2) Jerry Cruncher (3) Defarge (4) Barsad (5) Roger Cly 4

29. Defarge found the paper written by Dr. Manette (1) in Soho (2) at Tellson's (3) in the Bastille (4) in the wine shop (5) nowhere 3

30. The reading of the paper before the tribunal disclosed (1) the plan of the revolutionists (2) the names of the condemned (3) the resolution of Sydney Carton (4) Solomon's reason for changing his name (5) the secret of Madam Defarge's enmity toward Darnay 2✓

31. To make Charles Darnay unconscious so he could be removed from the Conciergerie, Sydney Carton used (1) a pistol (2) a club (3) his sword (4) wine (5) chemicals 5

32. Darnay was carried from prison upon the order of (1) Barsad (2) Evrémonde (3) Roger Cly (4) Defarge (5) nobody 1

33. The explosion of the pistol which killed Madam Defarge resulted also in (1) bringing aid (2) protecting Lucie (3) causing Miss Pross' deafness (4) alarming Jerry (5) setting the house on fire 3

34. Sydney Carton was accompanied to the guillotine by (1) Defarge (2) Darnay (3) a little seamstress (4) Lucie (5) an Irish policeman 3

35. Lucie's happiness was made secure by the (1) burial of the shoemaker's tools (2) death of Miss Pross (3) reading of the paper (4) sacrifice of Sydney Carton (5) murder of the Marquis 4

36. The plot of *A Tale of Two Cities* is (1) not important (2) slow-moving (3) carelessly constructed (4) difficult to follow (5) well developed 5

37. The lives of the leading characters become entangled in the scene in which (1) Dr. Manette is found (2) Carton takes the place of Darnay in prison (3) Darnay is arrested in France (4) Darnay is tried for treason (5) the wine is spilt 1√

38. An important turning point in the plot is the (1) death of Sydney Carton (2) receipt of the letter from France by Darnay (3) "fishing" trip of Jerry (4) finding of Miss Pross' brother (5) journey of Mr. Lorry to Dover 3√

39. One of the chief reasons for introducing Jerry Cruncher was to (1) testify to the identity of the English spy (2) deliver the message to Mr. Lorry (3) walk with Miss Pross (4) go fishing (5) show how children should be trained 1

40. The ignoble death of Madam Defarge is in contrast to the (1) death of the Marquis (2) murder of the child in the streets (3) death of the French king (4) death of Sydney Carton (5) death of Dr. Manette's patient 4

41. The catastrophe is the (1) marriage of Charles Darnay and Lucie Manette (2) imprisonment of Dr. Manette (3) murder of the Marquis (4) trial of Charles Darnay (5) death of Sydney Carton 5

Characters

Directions: Write the number of the best ending for each sentence given below on the marginal line.

1. Sydney Carton was (1) a leader in the French Revolution (2) a distinguished English lawyer (3) the driver on a coach (4) an English spy (5) the self-sacrificing lover of Lucie Manette 5
2. Charles Darnay was (1) a grave robber (2) a hired witness (3) the leader of the French aristocrats (4) Lucie Manette's husband (5) an Oxford professor of French 4
3. Dr. Manette was (1) a French doctor who had been unjustly imprisoned (2) a professional shoemaker (3) Lucie Manette's husband (4) an English country doctor (5) an eminent chemist 2√
4. Ernest Defarge was the (1) keeper of the Bastille (2) messenger of a bank (3) owner of a knitting mill (4) leader of the common people of Paris (5) nobleman from Paris 4
5. Mr. Lorry was (1) a French aristocrat (2) a trusted employee of Tellson's bank (3) the doctor of Beauvais (4) an English lawyer (5) Old Bailey 3√
6. Cruncher is the name of (1) a machine for crushing stone (2) a grave digger (3) an odd-job man for Tellson's (4) a flopper (5) an aggerawayter 3
7. Lucie Manette was (1) lovable (2) haughty (3) unsympathetic (4) self-sacrificing (5) illiterate 1
8. Madam Defarge was (1) sympathetic (2) demure (3) scurrilous (4) ruthless (5) sagacious 4
9. Miss Pross was (1) fascinating (2) heartless (3) eccentric (4) inefficient (5) effeminate 2√
10. The Marquis was (1) merciless (2) chivalrous (3) barbarous (4) sincere (5) compassionate 1
11. Mr. Stryver was (1) timid (2) self-conscious (3) ambitious (4) literary (5) sympathetic 3
12. The character that shows most development in the story is (1) Charles Darnay (2) Mr. Lorry (3) Defarge (4) Lucy Manette (5) Sydney Carton 4√
13. Madam Defarge may be said to represent the (1) justice of the Revolution (2) brute force of the Revolution (3) blind fury of the Revolution (4) intellect of the Revolution (5) faithful wife 4

14. Defarge may be said to represent the (1) lover of humanity (2) justice of the Revolution (3) brute force of the Revolution (4) upper class (5) faithful husband 3

15. Lucie Manette possessed (1) little personality (2) hatred for the French people (3) a desire for revenge (4) love for Sydney Carton (5) strong will power 1

16. Mr. Lorry may be said to be a (1) villain (2) balance wheel (3) highly intellectual man (4) foolish business man (5) grotesque character 2

17. An oddity in the story is (1) Carton (2) Darnay (3) Dr. Manette (4) Lucie Manette (5) Miss Pross 3✓

18. In contrast to Sydney Carton is (1) Darnay (2) Mr. Lorry (3) Stryver (4) Defarge (5) Dr. Manette 5✓

19. The Marquis (1) loved his nephew (2) had a distinct personality (3) sympathized with his servants (4) represented a type (5) was generous 3✓

20. Sydney Carton was transformed through (1) hard work (2) worry (3) drink (4) love (5) desire for revenge 4

Vocabulary

Directions: Write on the line at the left of each lettered term the number of the term defined.

1.	Old Bailey	*a.*	—— a general truth
2.	guillotine	*b.*	—— a blacksmith
3.	tumbril	*c.*	—— warlike
4.	pillory	*d.*	—— a criminal court in London
5.	aphorism	*e.*	—— selfish
6.	tribunal	*f.*	—— essential
7.	parricide	*g.*	—— a beheading machine
8.	farrier	*h.*	—— a device for punishment
9.	maligned	*i.*	—— a shield
10.	belligerent	*j.*	—— farmer's dump cart
11.	requisite	*k.*	—— a court of justice
12.	palpable	*l.*	—— injured
13.	lethargy	*m.*	—— one who murders a relative
14.	escutcheon	*n.*	—— drowsiness
15.	mercenary	*o.*	—— obvious

Key for Objective Test on Dickens' *A Tale of Two Cities*

Authors and His Works		Background and Setting	Plot		Characters	Vocabulary
1. 0	22. +	1. French Revolution	1. 1	22. 5	1. 5	*a.* 5
2. +	23. 0	2. Carlyle's *French Revolution*	2. 4	23. 2	2. 4	*b.* 8
3. +	24. 0	3. Paris	3. 3	24. 3	3. 1	*c.* 10
4. 0	25. 0	4. George III	4. 2	25. 3	4. 4	*d.* 1
5. +	26. 0	5. Revolutionary	5. 3	26. 3	5. 2	*e.* 15
6. 0	27. +	6. Louis XVI	6. 2	27. 1	6. 3	*f.* 11
7. 0	28. +	7. Marie Antoinette	7. 2	28. 4	7. 1	*g.* 2
8. +	29. +	8. Common people	8. 1	29. 3	8. 4	*h.* 4
9. +	30. 0	9. Goldsmith	9. 2	30. 5	9. 3	*i.* 14
10. 0	31. +	10. highwaymen (robbers)	10. 5	31. 5	10. 1	*j.* 3
11. +	32. 0	11. death (hanging)	11. 1	32. 1	11. 3	*k.* 6
12. 0	33. +	12. tragic	12. 1	33. 3	12. 5	*l.* 9
13. +	34. 0	13. Dover	13. 2	34. 3	13. 4	*m.* 7
14. +	35. +	14. autumn	14. 5	35. 4	14. 3	*n.* 13
15. +	36. +	15. misty	15. 4	36. 5	15. 1	*o.* 12
16. +	37. +	16. Bastille	16. 1	37. 4	16. 2	
17. +	38. 0	17. Bastille	17. 2	38. 2	17. 5	
18. 0	39. +	18. Paris (France)	18. 3	39. 1	18. 4	
19. +	40. 0	19. The Frozen Deep	19. 4	40. 4	19. 4	
20. 0	41. +		20. 1	41. 5	20. 4	
21. 0	42. 0		21. 2			

REFERENCES

Asher, W. — "The Reliability of Tests Requiring Alternative Responses," *Journal of Educational Research*, 9 : 234–240, March, 1924.

Bardy, J. H. — *An Investigation of the Written Examination as a Measure of Achievement with Particular Reference to General Science.* Philadelphia, 176 pp., 1923.

Barthelmess, H. M. — "Reply to a Criticism of Tests Requiring Alternative Responses," *Journal of Educational Research*, 6 : 357–359, Nov., 1922.

Batson, W. H. — "Reliability of the True-False Form of Examination," *Educational Administration and Supervision*, 10 : 95–102, Feb., 1924.

Brinkley, S. G. — *Values of New Type Examinations in the High School*, Teachers College Contributions to Education, Vol. 161, Bureau

of Publications, Teachers College, Columbia University, New York, 1924.

Butler, W. F. — "The Value of Informal Tests," *First Year Book, Elementary School Principals*, pp. 94–119, 1922.

Chapman, J. C. — "Individual Injustice and Guessing in the True-False Examination," *Journal of Applied Psychology*, 6 : 342–348, Dec., 1922.

DuBreuil, A. J. — "True-False Test in Literature and Formal English," *Bulletin of Illinois Association of Teachers of English*, May 1, 1923.

— "An Answer to Professor Leonard's Dangers of the True-False Test in English," *Bulletin of Illinois Association of Teachers of English*, April 1, 1924.

Eaton, M. P. — "New Style Examinations at Wadleigh High School," *New York Bulletin of High Points*, pp. 3–16, June, 1923.

Farwell, H. W. — "The New Type Examinations in Physics," *School and Society*, 19 : 315–322, March 15, 1924.

Gates, A. I. — "The True-False Test as a Measure of Achievement in College Courses," *Journal of Educational Research*, 12 : 276–287, May, 1921.

Geyer, D. L. — "A Uniform Objective Examination on Intelligence Testing," *Journal of Educational Psychology*, 15 : 372–375.

Glenn, E. R. — "The Conventional Examination in Chemistry and Physics *versus* the New Types of Tests," *School Science and Mathematics*, 21 : 666–670, Oct., 1921; 746–756, Nov., 1921; 23 : 459–470, May, 1923.

Hahn, H. H. — "A Criticism of Tests Requiring Alternative Responses," *Journal of Educational Research*, 6 : 236–240, Oct., 1922.

Hatch, R. W. — "How Can Examinations in History be Improved?" *Horace Mann Studies in Education*, Feb., 1922.

Holzinger, K. J. — "On Scoring Multiple Response Tests," *Journal of Educational Psychology*, 15 : 445–447, Oct., 1924.

Kimmel, W. C. — "Practice Tests in the Social Studies," *Historical Outlook*, 14 : 354–358, Dec., 1923.

Knight, F. B. — "Data on the True-False Test as a Device for College Examination," *Journal of Educational Psychology*, 13 : 75–80, Feb., 1922.

Kohs, S. C. — "High Test Scores by Sub-average Minds," *Psychological Bulletin*, 17 : 1–5, Jan., 1920.

Laird, D. A. — "A Comparison of the Essay and the Objective Type of Examinations," *Journal of Educational Psychology*, 14 : 123–124, Feb., 1923.

LAIRD, D. A. — "Notes on the Shortening of Examinations," *Journal of Educational Psychology*, 15 : 116–117, Feb., 1924.

McCALL, W. A. — "A New Kind of School Examination," *Journal of Educational Research*, 1 : 33–40, Jan., 1920.

— *How to Measure in Education*, ch. 4. The Macmillan Company, New York; 416 pp., 1922.

McCLUSKY, H. Y., and CURTIS, F. D. — "A Modified True-False Test," *Journal of Educational Research*, 14 : 213–224, Oct., 1926.

MELVIN, A. A. — "A True-False Test in English Literature," *English Journal*, 11 : 491–496, Oct., 1922.

MILLER, G. F. — "Formulas for Scoring Tests in which the Maximum Amount of Chance is Determined," *Journal of Educational Psychology*, 16 : 304–315, May, 1925.

MONROE, W. S. — "Examples of History Tests," *Historical Outlook*, 14 : 306–318, Nov., 1923.

— *An Introduction to the Theory of Educational Measurements*, ch. 11. Houghton Mifflin Company, Boston; 364 pp., 1923.

— *Written Examinations and Their Improvement*, ch. 3. *Ref.* Ch. 2.

MONROE, W. S., and SOUDERS, L. B. — "The Present Status of Written Examinations," *Ref.* Ch. 2.

MONROE, W. S., DEVOSS, J. C., and KELLY, F. J. — *Educational Tests and Measurements*, ch. 14. *Ref.* Ch. 2.

MOYER, F. E. — "New Types of History Tests," *Historical Outlook*, 14 : 323–324; Nov., 1923.

ODELL, C. W. — "Another Criticism of Tests Requiring Alternative Responses," *Journal of Educational Research*, 7 : 326–330, April, 1923.

PARKER, E. P. — "A Few Suggestions for Informal Testing in Geography," *Elementary School Journal*, 23 : 444–447, Feb., 1923.

PATERSON, D. G. — "Do New and Old Type Examinations Measure Different Mental Functions?" *School and Society*, 24 : 246–248, Aug. 21, 1926.

PATERSON, D. G., and LANGLIE, T. A. — "Empirical Data on the Scoring of True-False Tests," *Journal of Applied Psychology*, 9 : 339–348, Dec., 1925.

REMMERS, H. H., and Others — "An Experimental Study of the Relative Difficulty of True-False, Multiple Choice, and Incomplete-Sentence Types of Examination Questions," *Journal of Educational Research*, 14 : 366–371, Sept., 1923.

REMMERS, H. H., and REMMERS, E. M. — "The Negative Suggestion Effect on True-False Examination Questions," *Journal of Educational Psychology*, 17 : 52–56, June, 1926.

RUCH, G. M., and DEGRAFF, M. H. — "Corrections for Chance and 'Guess' *versus* 'Do Not Guess' Instructions in Multiple-Response Tests," *Journal of Educational Psychology*, 17 : 368–375, Sept., 1926.

RUCH, G. M. — *Improvement of the Written Examination. Ref.* Ch. 2.

RUCH, G. M., and STODDARD, G. D. — "Comparative Reliabilities of Five Types of Objective Examinations," *Journal of Educational Psychology*, 16 : 89–103, Feb., 1925.

RUSSELL, C. — *Classroom Tests.* Ginn and Company. Boston, Mass.; 346 pp., 1926.

SHRYOCH, R. H. — "New Tests for Old," *Historical Outlook*, 14 : 319–322, Nov., 1923.

SPENCER, P. L. — "The Improvement of Teaching by Means of 'Home Made' Non-standardized Diagnostic Tests and Remedial Instruction," *School Review*, 31 : 276–281, April, 1923.

THORNDIKE, E. L. — *New Methods in Arithmetic*, ch. 13. Rand, McNally & Co., Chicago; 260 pp., 1921.

— *Psychology of Algebra*, ch. 6. The Macmillan Company, New York; 483 pp., 1923.

TOOPS, H. A. — *Trade Tests in Education*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 115; Bureau of Publications, Teachers College, Columbia University, New York; 118 pp., 1921.

WEIDEMANN, C. C. — *How to Construct the True-False Examination*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 225; Bureau of Publications, Teachers College, Columbia University, New York; 118 pp., 1926.

WEST, P. V. — "A Critical Study of the Right Minus Wrong Method," *Journal of Educational Research*, 8 : 1–9, June, 1923.

WILKINS, L. A. — "Suggestions as to the Formulation of Questions in Standardized Examination in Modern Languages," *New York Bulletin of High Points*, pp. 27–36, June, 1923.

WOOD, B. D. — *Measurement in Higher Education*, chs. 7–12. *Ref.* Ch. 2.

— "Measurement of Law School Work," *Columbia Law Review, Vol. 24, No. 3*, March, 1924; and *Vol. 25*, March, 1925.

— "Studies of Achievement Tests," *Journal of Educational Psychology*, 17 : 1–22, 125–139, 263–269, Jan., Feb., and April, 1926.

WOODY, C. — "Informal Tests as a Means for the Improvement of Instruction," *First Yearbook, Elementary School Principals*, pp. 87–94, 1922.

CHAPTER IV

INTELLIGENCE TESTS FOR USE IN HIGH SCHOOL

History of Intelligence Testing

Intelligence testing is the resultant of at least five converging movements, some practical, some theoretical. Perhaps foremost of the movements was the very practical one of dealing with the feeble-minded. Their treatment is distinctly a medical and educational problem, but the diagnosis of feeble-mindedness in its different degrees leads at once to the problem of mental measurement or diagnosis. It is perhaps unfair to ascribe the beginnings of this movement to Binet, but what was done before his time seems insignificant beside his contribution.[1] Binet, a Frenchman (1857–1911), the son of a physician, was a genius whose interests were both theoretical and practical. His period of production consisted of vague gropings in the field of applied psychology, usually in the direction of diagnosing feeble-mindedness and of the treatment of special classes in school. His investigations, driven by an unquenchable spirit, and guided by a great intellect, issued in 1905 in a set of tests known as the 1905 Binet scale. These tests were a set of tasks to be performed under controlled conditions and the responses were more or less defined. In 1908 the first scale appeared with tests grouped in age levels and the mental

[1] Binet, A., and Simon, Th. — "Sur la nécessité d'établir un diagnostic scientifique des états inferieurs de l'intelligence"; "Methods nouvelles pour le diagnostic du niveau intellectuel des anormaux"; "Application des méthods nouvelles au diagnostic du niveau intellectuel chez des enfants anormaux et anormaux d'hospice et d'école primaire"; *L'Année Psychologique*, Vol. 11, 1905.

diagnosis was given in terms of mental age. A revision, essentially the series of tests used to-day, appeared in 1911. Since that time many revisions have appeared abroad and in America. The scale is being constantly refined, but in principle it is the same as that originally proposed by Binet. In 1896 the University of Pennsylvania established a psychological clinic. Since then such clinics have been founded in universities, city school systems, and hospitals; and the method of individual mental testing has been largely employed.

A second movement which has helped intelligence testing is the movement of experimental psychology. The beginning of this movement may be dated at 1879, when Wundt established his laboratory in Leipzig. One of our most useful mental tests — the completion test — was developed in this movement by Ebbinghaus [1] in 1897.

A third line of attack is the study of individual differences. Cattell [2] in 1890 wrote on "Mental Tests and Measurements" and urged standardization of procedure and the establishment of norms. Cattell himself in the next decade did much to develop mental testing in his work in Columbia College. In 1896 Cattell and Farrand [3] published the first results of these tests. This movement had a consistent and steady growth, being stimulated largely through the efforts of Thorndike [4] of Columbia.

[1] EBBINGHAUS, H. — "Über eine neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bei Schulkindern," *Zeitschrift fur angewandte Psychologie*, vol. 13, 1897.

[2] CATTELL, J. McK., and FARRAND, L. — "Mental Tests and Measurements," *Mind*, 15: 373–380, 1890.

[3] CATTELL, J. McK., and FARRAND, L. — "Physical and Mental Measurements of the Students of Columbia University," *Psychological Review*, 3: 618–648, 1896.

[4] THORNDIKE, E. L. — *An Introduction to the Theory of Mental and Social Measurements*, 1904.

Two other movements have been contributory. One, which is fourth in our list, is the Galton[1]-Pearson[2] movement of Great Britain. The contribution from this school is largely one of statistical technique. The English have on the whole given little in the way of new testing material, but they have made an incomparable contribution in the exact methods of building up tests and treating the results.

The last movement is that contributed by the anthropologists who have endeavored to "rate the relative intelligence of various races and to scale men of talent and genius according to scientific standards." Names to be mentioned here are Galton, Cattell, Boas, Ellis, and Woods.[3]

The Binet test is an individual test and can properly be given to only one child at a time. Another great advance in mental testing was the testing of several individuals at one sitting. This group test movement bids fair to usurp the field of mental testing, converging as it does all the streams of the testing movement. It uses materials originated by the experimentalist, the individual difference investigators, and the clinicians, and the methods and technique of the English Galton-Pearson school. Pyle[4] gave several tests to groups of children, the results of which were published in 1913. Thorndike used a group mental test in the examination of employees of the Metropolitan Life Insurance Company in 1914, of which no results were published.

[1] GALTON, F. — *Hereditary Genius*, 1869.

GALTON, F. — *Inquiries into Human Faculty*, 1883.

[2] PEARSON, K. — "Skew Variation in Homogeneous Material," *Phil. Tran. A.*, 186: 343 ff., 1895.

PEARSON, K. — "Regression, Hereditary and Panmixia," *Phil. Tran. A.*, 187: 253–318, 1896.

[3] GALTON, F. — *Op. cit.*

CATTELL, J. McK. — *The Mind of Primitive Man*, 1911.

ELLIS, H. — *Study of British Genius*, 1904.

WOODS, F. A. — *Mental and Moral Heredity in Royalty*, 1906.

[4] PYLE, W. H. — *The Examination of School Children*, 1913.

Pintner[1] in 1917 used a set of six tests on a group of children. Miller[2] sectioned pupils in the University of Minnesota High School on the basis of mental tests in 1917. To Otis,[3] however, properly belongs the credit for compiling and publishing in 1918 the first group test of intelligence as a measuring instrument.

It is now history how the World War gave impetus to the movement in the United States. When the need for sorting out the draftees became imperative, the government appealed to the psychologists of the nation for aid. A Committee on the Psychological Examination of Recruits was formed composed of Yerkes (chairman), Bingham, Goddard, Haines, Terman, Wells, and Whipple. "The group tests in manuscript already devised by Otis were made the basis of much of the work of this committee. Four of the ten tests in the original Army scale for group testing were taken from the Otis scale practically without change. . . . Between September, 1917, and January, 1919, about 1,727,000 men were examined in the Army by means of the psychological tests devised for this purpose."[4] These army tests were so useful that during and after the war there was a movement to make them do for the schools in sorting children what they had done for the Army in sorting recruits. The Otis test was pressed into immediate service in the secondary school. The "Army Alpha" test itself has been given very widely in high schools. In 1920 Terman published his "Group Test of Mental Ability." This test has

[1] PINTNER, R. — "The Mentality of the Dependent Child Together with a Plan for a Mental Survey of an Institution," *Journal of Educational Psychology*, 8: 76–88, April, 1917.

[2] MILLER, W. S. — "The Administrative Use of Intelligence Tests in the High School," *Twenty-first Yearbook of the National Society for the Study of Education*, p. 204.

[3] OTIS, A. S. — "An Absolute Point Scale for the Group Measure of Intelligence," *Journal of Educational Psychology*, 9: 239–261; 333–348, May, June, 1918.

[4] WYLIE, A. T. — "A Brief History of Mental Tests," *Teachers College Record*, 23: 19–33, Jan., 1922.

had great popularity because of its careful construction. Miller's "Mental Ability Test" was first published in 1921 and takes place with the other tests mentioned for use in high school.

This survey of the group testing movement as applied to high schools should not omit the tests in a higher level constructed by Thorndike and Thurstone in 1919. Thorndike's "Intelligence Examination for High School Graduates" was compiled in 1919 for use in selecting students for admission to Columbia College. Thurstone's "Psychological Examination for College Freshmen and High School Seniors" was constructed with the same function in mind. All these recent tests have gathered together and refined the hopeful results of the different movements. The refining process is still going on.

Perhaps the more skeptical reader will have been asking before now what all this is about. What is meant by an intelligence test? What is intelligence? Let us review briefly the situation. Practically, the need for mental tests arose in connection with the diagnosis of the feeble-minded. When society took upon itself the care of those unfortunate ones who did not have the nervous organization to care for themselves, it was brought face to face with the necessity of setting a definition to feeble-mindedness and of devising methods of differentiating between those who do and those who do not come under this definition. Likewise wherever it is necessary to choose between men (or children) to decide who shall be promoted in school, who shall be sent to a higher job, who shall have the salary raised, who shall become officers in the Army, society welcomes any objective means of making the differentiation. The development of the Army tests was conditioned wholly by practical exigencies. The use of mental tests in schools since the war, which promises to become permanent, likewise was occasioned by the desire of teachers for some objective means of classifying and sectioning children, which would supersede their faulty subjective methods. So the mental testing movement

was born of a practical need and not of any theoretical attempt to measure "native intelligence."

Yet these same tests which are so useful in sorting and classifying do, at the same time, measure intelligence, and that is why they are boldly titled "Mental Tests" or "Intelligence Tests." Perhaps it would be better for the movement if they had been called "Classification Tests" or some similar title. At any rate, under such a title they would still perform their function as well and would not lay themselves open to such violent attacks as have been given by Lippman,[1] Bagley,[2] and others. The evidence that these tests are tests of native intelligence may be found in Freeman's *Mental Tests* or Thorndike's *Measurement of Intelligence.*

Description of the Tests

OTIS GROUP INTELLIGENCE SCALES

Otis' original test[3] is now one of our best tests for use in high school. This test is called the "Otis Group Intelligence Scale," constructed by Arthur S. Otis. There are two scales, one a Primary Examination and the other an Advanced Examination. The Advanced Examination is the one suitable for use in grades five to twelve, which includes the high school. The test comes in two forms, identical in structure but containing different material. Most of the best tests to-day come in two or more forms, equivalent in difficulty throughout, so that they may be more generally useful in case one wishes to repeat the test. Since it was the first of the published group tests of

[1] Lippman, W. — "The Mental Age of Americans," *New Republic*, 32: 213–215, Oct. 25; 246–248, Nov. 1; 275–277, Nov. 8; 297–298, Nov. 15; 328–330, Nov. 22; 33: 9–10, Nov. 29, 1922.

[2] Bagley, W. C. — "Educational Determinism; or Democracy and the I. Q.," *School and Society*, 15: 373-384, April, 1922.

[3] Published by the World Book Company.

intelligence it has served as a model in subsequent work and receives an extended description here.

The tests are as follows:

TABLE 5

OTIS GROUP INTELLIGENCE SCALE, ADVANCED EXAMINATION TESTS, TIME ALLOWANCE, AND POSSIBLE SCORE

	Test	Time in Minutes	Total Score Possible
1	Following Directions	5	20
2	Opposites	$1\frac{1}{2}$	25
3	Disarranged Sentences	$1\frac{1}{2}$	25
4	Proverbs	6	20
5	Arithmetic	6	20
6	Geometric Figures	6	20
7	Analogies	3	25
8	Similarities Test	4	20
9	Narrative Completion	6	25
10	Memory	$6\frac{1}{2}$ *	30
		$45\frac{1}{2}$	230

* $3\frac{1}{2}$ minutes for reading the story and 3 minutes for writing.

A brief description of each test is afforded by the directions.[1]

Test 1. Following Directions.

Sample problem: Write the fifth letter of the alphabet. ( )

Test 2. Opposites

Directions: Look at the first word in each line, think what word means exactly the opposite of it, find that word among the five words in parenthesis in that line and draw a line under it.

up (short, <u>down</u>, small, low, young)
hot (warm, ice, dark, <u>cold</u>, fire)

[1] From Otis Group Intelligence Scale, Advanced Examination. Copyright 1918 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

Test 3. Disarranged Sentences.

Directions: The words on each line below make one sentence if put in order. If the sentence the words make is *true*, underline the word *true* at the side of the page. If the sentence they would make is *false*, underline the word *false*.

men money for work	(true false)
uphill rivers flow all	(true false)
ocean waves the has	(true false)

Test 4. Proverbs.

Directions: Read each proverb, find the statement that explains it, and put the number of that statement in the parenthesis before the proverb.

Proverbs.

(3) The burnt child fears the fire.
( ) Rome was not built in a day.
( ) There is no smoke without fire.
Etc.

Statements.

1. Time is required to produce anything of value.
2. Failure follows frequent change of plan.
3. Unhappy experiences teach us to be careful.
4. Those in disgrace always want to disgrace others.
5. There is no result without a cause.

Etc.

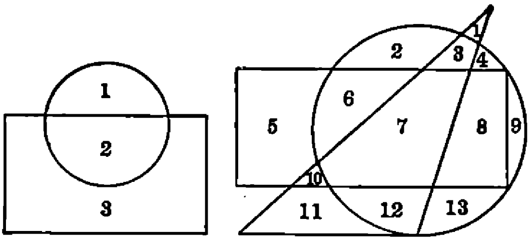
Test 5. Arithmetic.

Directions: Place the answer to each problem in the parenthesis after the problem. Do any figuring you wish on the margin of the page.

1. If a boy had 10 cents and earned 5 cents, how much money did he have then? ( ) cents
10. If a man walks east from his home 7 blocks and then walks west 4 blocks, how far is he from home? ( ) blocks
20. If a wire 20 inches long is to be cut so that one piece is $\frac{2}{3}$ as long as the other piece, how long must the original piece be? ( ) inches

Test 6. Geometric Figures.

Directions: Each problem asks a question that is answered by a number. Write the answer to each problem in the parenthesis after the statement of the problem.



Look at the figure. What number is in the circle but not in the rectangle? (1)

Test 7. Analogies.

Directions: The first sample means: Finger is to hand as toe is to what? Underline the word on each line that should go in the parenthesis in the place of the question mark.

finger : hand — toe : (?) — foot, knee, arm, shoe, nail.

Test 8. Similarities Test.

Directions: Find the way in which the first three things on a line are alike. Then look at the five other things on the same line and draw a line under the one that is *most* like the first three.

hat, collar, glove . . . . hand, cane, head, shoe, house
rose, daisy, violet . . . . bush, red, plant, bed, pansy
desk, bed, chair . . . . book, table, floor, pencil, coat

Test 9. Narrative Completion.

(The test consists of a story with words frequently deleted. These words may be filled in by choosing words from columns placed at the right of the story.)

Directions: For each numbered blank in the story, choose the best word of the three in the list having the same number as the blank. Underline the word you choose. You may write these words in the blank spaces if you wish but only the underlining counts. Do nothing about the blanks that are not numbered.

The Reward of Kindness

Once upon a —— (1) there was a —— (2) that lived in a —— (3) etc.

Underline words here.

1. time place man
2. man lion dog
3. street garden forest

Test 10. Memory.

(First a story is read to the class. Then the class answers from memory questions about the story.)

Directions: Read each question and if the right answer according to the story is *yes*, draw a line under the word *yes*. If the right answer is *no*, draw a line under the word *no*. But if you do not know the right answer, because the story didn't say, draw a line under the words *didn't say*.

Was the story about a king?	(yes no didn't say)
Was the king's daughter 16 years old?	(yes no didn't say)
Was she ugly?	(yes no didn't say)

In order to give the test one must have a copy of the manual of directions, which must be ordered separately. One feature of a standardized test is that the whole procedure is standardized. In the Otis examination, for instance, every word and act of the examiner from the time he enters the room until he leaves is definitely prescribed in the manual of directions. The instructions for giving the test which the examiner reads aloud are printed in heavy type in the manual and should be followed exactly with no deviations.

For scoring the test a key is provided with the correct answers which may be placed directly beside the pupil's answers. In this way scoring is facilitated. In all cases the score is the number of correct elements, except in the case of test number 3, when the right minus wrong method described in chapter 3 is employed. The scores on each test are entered at the bottom of the page and later the scores on all the separate tests are transferred to the cover page and totaled.

The norms for the Otis Advanced Examination are given in the following tables.

In Table 6 grade norms are given and in Table 7 age norms. After age 16 there is very little increase in the age norms. At one time it was thought that mental growth stopped at or before 16, but later investigations seem to show that pupils who

stay on through high school continue their growth as measured by a test like the Otis test.

TABLE 6

OTIS GROUP INTELLIGENCE SCALE — ADVANCED EXAMINATION
GRADE NORMS

GRADE	10%	25%	50%	75%	90%
4	19	29	44	62	80
5	36	48	63	80	97
6	55	69	85	102	117
7	65	80	97	113	128
8	80	95	112	128	144
9	92	108	125	142	157
10	106	120	136	153	166
11	113	131	149	167	184
12	120	136	154	169	183

TABLE 7

OTIS GROUP INTELLIGENCE SCALE — ADVANCED EXAMINATION
AGE NORMS

YEAR	8	9	10	11	12	13	14	15	16	17	18	19
MONTH												
0	25	40	55	68	80	90	100	110	120	127	130	130
1	26	41	56	69	81	91	101	111	121	127	130	130
2	27	42	57	70	82	92	102	112	122	128	130	130
3	28	43	58	71	83	93	103	113	123	128	130	130
4	30	45	59	72	84	94	104	114	123	128	130	130
5	31	46	60	73	85	95	105	115	124	128	130	130
6	32	47	61	74	85	95	105	115	124	129	130	130
7	33	48	62	75	86	96	106	116	125	129	130	130
8	35	50	63	76	87	97	107	117	125	129	130	130
9	36	51	64	77	88	98	108	118	126	129	130	130
10	37	52	65	78	89	99	109	119	126	129	130	130
11	38	53	66	79	90	100	110	120	127	129	130	130

TABLE 8

OTIS GROUP INTELLIGENCE SCALE — ADVANCED EXAMINATION MENTAL AGE NORMS

Tens of Score	20	30	40	50	60	70	80	90	100	110
Units of Score										
0		8 : 4	9 : 0	9 : 8	10 : 5	11 : 2	12 : 0	13 : 0	13 : 11	14 : 10
1		8 : 5	9 : 1	9 : 9	10 : 6	11 : 3	12 : 1	13 : 1	14 : 0	14 : 11
2		8 : 6	9 : 2	9 : 10	10 : 7	11 : 4	12 : 2	13 : 2	14 : 1	15 : 0
3		8 : 7	9 : 3	9 : 11	10 : 8	11 : 5	12 : 3	13 : 3	14 : 2	15 : 1
4		8 : 7	9 : 3	9 : 11	10 : 9	11 : 6	12 : 4	13 : 4	14 : 3	15 : 2
5	8 : 0	8 : 8	9 : 4	10 : 0	10 : 10	11 : 7	12 : 6	13 : 6	14 : 5	15 : 3
6	8 : 1	8 : 9	9 : 5	10 : 1	10 : 11	11 : 8	12 : 7	13 : 7	14 : 6	15 : 4
7	8 : 2	8 : 10	9 : 6	10 : 2	10 : 11	11 : 9	12 : 8	13 : 8	14 : 7	15 : 5
8	8 : 3	8 : 11	9 : 7	10 : 3	11 : 0	11 : 10	12 : 9	13 : 9	14 : 8	15 : 6
9	8 : 3	8 : 11	9 : 7	10 : 4	11 : 1	11 : 11	12 : 10	13 : 10	14 : 9	15 : 7

Tens of Score	120	130	140	150	160	170	180	190	200
Units of Score									
0	15 : 8	16 : 4	17 : 0	17 : 6	18 : 0	18 : 6	18 : 11	19 : 2	19 : 6
1	15 : 8	16 : 5	17 : 1	17 : 7	18 : 1	18 : 6	18 : 11	19 : 3	
2	15 : 9	16 : 6	17 : 1	17 : 7	18 : 1	18 : 7	19 : 0	19 : 3	
3	15 : 10	16 : 7	17 : 2	17 : 8	18 : 2	18 : 8	19 : 0	19 : 3	
4	15 : 11	16 : 8	17 : 3	17 : 9	18 : 3	18 : 8	19 : 0	19 : 4	
5	16 : 0	16 : 8	17 : 3	17 : 9	18 : 3	18 : 9	19 : 1	19 : 4	
6	16 : 1	16 : 9	17 : 4	17 : 10	18 : 4	18 : 9	19 : 1	19 : 4	
7	16 : 2	16 : 10	17 : 5	17 : 11	18 : 4	18 : 10	19 : 1	19 : 5	
8	16 : 3	16 : 11	17 : 5	17 : 11	18 : 5	18 : 10	19 : 2	19 : 5	
9	16 : 4	16 : 11	17 : 6	18 : 0	18 : 5	18 : 11	19 : 2	19 : 5	

OTIS SELF-ADMINISTERING TEST OF MENTAL ABILITY

More recently Otis has published a simplified test called the Otis " Self-Administering Test of Mental Ability." The Higher Examination is designed for grades 9 and higher. This examination is issued in alternate forms, *A* and *B*, alike except in content, and interchangeable in use. The advantage of this test is that it may be very easily given and scored. A separate manual of directions comes with the tests and should be used in administering them. Pupils are given time to read through and study the first page after the papers have been distributed. When practically all have read the first page, the class is allowed to turn the page and start the test. They are given 30 minutes without interruption to do as much as possible. All answers are given in figures written in parentheses at the right of each page. The scoring is easily done by comparing the answers with the key which is printed around the margin of the directions sheet.

The examination starts in as follows:

1. The opposite of hate is (?)
   1 enemy, 2 fear, 3 love, 4 friend, 5 joy ( )
2. If 3 pencils cost 5 cents, how many pencils can be bought for 50 cents? ( )
3. A bird does not always have (?)
   1 wings, 2 eyes, 3 feathers, 4 a nest, 5 a bill ( )

The Advanced Examination, previously described, was freely drawn upon for items to make up the Self-Administering Tests.

Form *B* is more difficult than form *A* on the average by 4 points. However, in repeating an intelligence test, even a different form of the same test, there is usually an increase in score. If form *B* is given the day following form *A*, the practice effect is practically counterbalanced by the 4 points increase in difficulty, so that under these conditions the two tests may be considered identical.

Following are norms:

TABLE 9

OTIS SELF-ADMINISTERING TEST OF MENTAL ABILITY, HIGHER EXAMINATION

GRADE NORMS

Grade	8	9	10	11	12
Norm	30	34	37	40	42

TABLE 10

OTIS SELF-ADMINISTERING TEST OF MENTAL ABILITY, HIGHER EXAMINATION

AGE NORMS

YEARS	12	13	14	15	16	17	18
MONTHS							
0	23	28	32	36	39	41	42
1	24	28	32	36	39	41	
2	24	29	33	37	39	41	
3	25	29	33	37	40	41	
4	25	29	33	37	40	41	
5	25	30	34	37	40	41	
6	26	30	34	38	40	42	
7	26	30	34	38	40	42	
8	27	31	35	38	40	42	
9	27	31	35	38	41	42	
10	27	31	35	39	41	42	
11	28	32	36	39	41	42	

The total score possible is 75 and the time allowance is thirty minutes, although this time may be shortened to twenty minutes, if desired. In the latter case the above norms do not apply.

ARMY ALPHA

The "Army Alpha" test[1] is the test that was so widely given during the war to the National Army. As previously stated, it

[1] Army Alpha may now be purchased from Division of Psychology, Surgeon General's Office, War Department, Washington, D. C.; or Kansas State Normal School, Emporia, Kansas; or C. H. Stoelting Company, 3037-3047 Carroll Ave., Chicago, Illinois.

was compiled, with Otis' manuscript as a basis, from the combined experience of such men as Angell, Dodge, Scott, Strong, Thorndike, Watson, and Woodworth. A preliminary form called "Army *a*" was used for try-outs and the eight most successful tests out of the ten in "Army *a*" were used to make "Army Alpha."

The directions for giving are also to be found in Yoakum and Yerkes *Army Mental Tests* and in VOLUME 15 of the NATIONAL ACADEMY OF SCIENCES PUBLICATIONS entitled *Psychological Examining in the United States Army*, on pages 157 to 161.

"Army Alpha" was made in five different forms, 5, 6, 7, 8, and 9. These are not of the same difficulty, as is shown in the following table (abridged from table 177 of Vol. 15 cited above):

TABLE 11

DIFFERENCE BETWEEN MEAN SCORES, ALPHA FORMS

FORM	DIFFERENCE BETWEEN MEANS	FORM	DIFFERENCE BETWEEN MEANS
5–6	6	6–8	−1
5–7	2	6–9	0
5–8	5	7–8	3
5–9	6	7–9	3
6–7	−4	8–9	1

Form 5 proves to be the easiest and to give the highest scores and form 6 the hardest and to give the lowest scores.

TABLE 12

ARMY ALPHA: TESTS, TIME ALLOWANCE, AND POSSIBLE SCORE

TESTS	TIME IN MINUTES	TOTAL SCORE POSSIBLE
1. Oral Directions	indefinite	12
2. Arithmetical Problems	5	20
3. Practical Judgments	1½	16
4. Synonym-Antonym	1½	40
5. Disarranged Sentences	2	24
6. Number Series Completion	3	20
7. Analogies	3	40
8. Information	4	40
	20 +	212

Norms for "Army Alpha" are given in the two following tables.

TABLE 13

ARMY ALPHA

GRADE NORMS

Grade	9	10	11	12
Score	96	111	123	126

TABLE 14

ARMY ALPHA

MENTAL AGE NORMS

MEDIAN MENTAL AGE	MEDIAN ALPHA SCORE
9–6	10
10–6	25
11–6	40
12–6	55
13–6	70
14–6	85
15–6	100
16–6	115
17–6	130
18–6	145
19–6	160
20–6	175

TERMAN GROUP TEST OF MENTAL ABILITY

This test,[1] which was an outgrowth of experience in the Army testing, appeared in 1920. It was constructed by Lewis M. Terman of Stanford University. The test is designed particularly

[1] Published by the World Book Company.

for use in high school and has proved its usefulness. It was carefully made. Thirteen tests were originally tried out, but only ten were retained for the final test. Out of 610 items which were tried only 370 were retained. These have been separated to make two forms, *A* and *B*, of approximately equal difficulty. "All items which failed to differentiate pupils of known brightness from pupils of known dullness were eliminated."[1] A manual of directions accompanies the tests. The directions in the manual should be followed verbatim.

The tests included are:

TABLE 15

TERMAN GROUP TEST OF MENTAL ABILITY: TESTS, TIME ALLOWANCE, AND POSSIBLE SCORE

Test	Time in Minutes	Total Possible Score
1. Information . . .	2	20
2. Best Answer	2	22
3. Word Meaning	2	30
4. Logical Selection . . .	3	20
5. Arithmetic	4	24
6. Sentence Meaning . . .	2	24
7. Analogies	2	20
8. Mixed Sentences . . .	3	18
9. Classification	3	18
10. Number Series	4	24
	27	220

Norms which are given in the manual of directions are repeated in Table 16. These norms represent the results of testing over 40,000 pupils in grades 7 to 12. The scores which are equaled or exceeded by 50 per cent of the pupils in any grade

[1] *Manual of Directions*, Terman Group Test of Mental Ability, p. 3. Copyright 1920 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

are known as *medians*, and may be taken as average scores made by each of the grades.

TABLE 16

TERMAN GROUP TEST OF MENTAL ABILITY: PERCENTILE NORMS BY GRADE

PER CENT	7	8	9	10	11	12
1 per cent equal or exceed	147	170	181	194	203	207
2½	134	159	172	185	196	200
5	122	148	164	177	189	194
10	109	135	151	166	180	185
15	100	126	142	159	174	179
20	93	118	135	152	168	174
25	88	112	128	147	163	169
30	83	107	123	141	158	165
40	75	97	113	131	147	156
50	68	89	104	122	138	147
60	61	81	95	113	128	138
70	54	73	86	103	118	128
75	51	69	81	98	112	122
80	47	64	76	92	105	115
85	43	58	71	86	99	109
90	38	52	63	79	90	100
95	31	43	53	67	77	86
97½	25	36	44	58	66	74
99	20	30	35	48	55	63

Table 17 gives mental age equivalents in terms of the Stanford-Binet Scale. These equivalents were determined by testing 306 pupils with both the Terman Group Test and the Stanford-Binet Scale. Terman notes that the mental age equivalents given in the table may be a little too high because the children who were tested had become test wise.

Terman says regarding his test that its purpose is to make a difference in the educational treatment of pupils and not to gratify a merely idle curiosity regarding their intellectual status.

TABLE 17

TERMAN GROUP TEST OF MENTAL ABILITY: MENTAL AGE EQUIVALENTS

Terman Group Test Score	Stanford-Binet Mental Age	Terman Group Test Score	Stanford-Binet Mental Age
210	19–6	110	14–10
205	19–3	105	14–7
200	19–0	100	14–4
195	18–9	95	14–1
190	18–7	90	13–11
185	18–4	85	13–8
180	18–1	80	13–5
175	17–10	75	13–2
170	17–7	70	12–11
165	17–5	65	12–9
160	17–2	60	12–6
155	16–11	55	12–3
150	16–8	50	12–0
145	16–5	45	11–9
140	16–3	40	11–6
135	16–0	35	11–4
130	15–9	30	11–1
125	15–6	25	10–10
120	15–3	20	10–7
115	15–1	15	10–4

MILLER MENTAL ABILITY TEST

W. S. Miller[1] of the University of Minnesota states that the inception of this test was in 1914, when, under the direction of Dr. Whipple, he prepared a thesis which looked definitely forward to the preparation of a battery of tests useful for the group-testing of high school students. Such testing was first carried on in 1917

[1] *Op. cit.*, p. 189.

in the University of Minnesota High School, but it was not until 1921 that the test appeared publicly from the press.[1] The Miller "Mental Ability Test" takes rank with the others mentioned as a good high school intelligence test. It now comes in two forms, *A* and *B*, although Miller has recently published data showing that form *B* is as much as 14 months more difficult than form *A*. The manual says that this difference may be compensated for by practice effect, but that is doubtful. This test is useful in the seventh grade and above.

This examination consists of three sub-tests:

TABLE 18

MILLER MENTAL ABILITY TEST: TESTS, TIME ALLOWANCE, AND POSSIBLE SCORE

Test	Time in Minutes	Total Score
I Disarranged Sentences .	8	40
II Word Meaning . . .	5	40
III Analogies	6	40
	19	120

The directions occupy the first page of the test booklet, but the examiner should provide himself with the extra booklet of manual of directions and should follow these directions closely.

The norms given in the manual of directions are repeated here.

TABLE 19

MILLER MENTAL ABILITY TEST: GRADE NORMS

Grade	7	8	9	10	11	12	13
Score	35	46	53	62	69	74	85

[1] Published by the World Book Company.

TABLE 20

MILLER MENTAL ABILITY TEST: PERCENTILE GRADE NORMS

GRADE	PERCENTILE SCORES										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
7	1	17	23	27	31	35	39	44	50	58	85
8	5	25	32	37	42	46	51	56	62	68	96
9	1	29	37	43	48	53	57	63	69	77	103
10	7	39	47	53	58	62	66	70	76	84	106
11	16	44	52	59	64	69	74	78	83	89	110
12	8	50	58	64	69	74	78	83	87	93	116
13	22	60	69	75	81	85	89	93	98	103	117

All of these tests will give satisfactory results in the high school. The Terman test has consistently shown itself superior, although the margin of superiority over the other tests is small. The Otis Advanced Examination is very satisfactory, although the completion test (No. 9) and the memory test (No. 10) have not been used in subsequent tests and appear to be slightly less useful. No. 10 is subject to subjective variations as it depends on the oral reading of a story by the examiner. The Terman test, although the best *for high school use,* is not difficult enough in its upper levels for superior pupils or for superior schools. The Army test designed for use on men in the army is naturally less well suited to school children. It contains many allusions to military events and conditions. On the face of it this test ought to be more difficult for girls than for boys for this reason, although the difference is slight.[1] The information test in "Army Alpha"

[1] Madsen and Sylvester report an average difference of about six points in high schools in favor of boys; Wentworth finds a superiority of 8.7 points, in comparing high school senior boys and girls. MADSEN, I. M., and SYLVESTER, R. H. — "High School Students' Intelligence Ratings according to the Army Alpha Test," *School and Society,* 10: 407–419, Oct. 4, 1919. WENTWORTH, M. M. — "Army Alpha Tests

(No. 8) contains allusions to advertisements, motion picture actors, and the like in vogue during the war period. As these become less well known this test is certain to become more difficult and less valid. Yet, notwithstanding these points, "Army Alpha" has been a very useful instrument for use in high schools. The Miller test can be given in the shortest time and it is remarkable that it stands up so well beside the longer tests. If a short test is wanted, use the Miller or the Otis Self-Administering Test.

THORNDIKE INTELLIGENCE TEST FOR HIGH SCHOOL GRADUATES

This test[1] cannot be described here in great detail. The test was first published in 1919. Thorndike prepared a large amount of material so that new editions are printed each year for college entrance purposes. The examination consists of four parts. Two of the parts are of the same type as the "Army Alpha," but more extensive and more difficult. The other two parts are of a different nature, including material found to measure intelligence well in its upper levels and to predict ability to do college work. "One test in the third part is inserted for the particular purpose of doing justice to boys whose intelligence and experience have been directed toward things rather than books." A unique feature of the test is the practice form. The purpose of the practice form is to provide familiarity with the nature of the test with the hope that thereby the practice effect will be lessened. For ten minutes the pupils are allowed to inspect this practice form, being told at intervals of a minute or so, according to a schedule, to turn from one test to another without doing any writing. At the end of the ten minutes it is collected.

and Teachers' Estimates in Hollywood High School," *School and Society*, 12: 58–60, July 10, 1920.

[1] Published by Bureau of Publications, Teachers College, Columbia University.

TABLE 21

THORNDIKE INTELLIGENCE EXAMINATION FOR HIGH SCHOOL GRADUATES: TESTS AND TIME ALLOWANCE

TEST	TIME IN MINUTES
Practice Form	10
Part I, Trial 1	30
Part I, Trial 2	30
Part II	60
Part III	40
	2 hours 50 minutes

From twenty to thirty-five minutes more must be allowed for explanation, distribution of papers, and rest period. Three and one half hours in all should be allowed for giving the test. The total number of credits is multiplied by 0.2 to turn them into scores comparable with the norms given.

TABLE 22

THORNDIKE INTELLIGENCE EXAMINATIONS FOR HIGH SCHOOL GRADUATES: FREQUENCIES OF SCORES IN PER CENTS

SCORES	FRESHMEN IN COLLEGES OF HIGH STANDARDS	FRESHMEN IN COLLEGES AND AGRICULTURAL SCHOOLS OF LOWER STANDARDS	WOMEN: LIBERAL ARTS EASTERN STATES	WOMEN: LIBERAL ARTS SOUTHERN STATES	NORMAL SCHOOL FIRST-YEAR STUDENTS EASTERN STATES
10–19	—	1½	—	—	—
20–29	—	6	—	—	½
30–39	1	14½	2	2	1½
40–49	3	24½	4	8	13
50–59	9	24	8	20	24
60–69	22	18½	28	34	29
70–79	26	8	30	21	21
80–89	21	1	20	10	10
90–99	13	2	8	5	1
100	5	—	—	—	—

A new series of the Thorndike Examination is being issued, starting with 1925, for which new norms will be required.

THURSTONE PSYCHOLOGICAL EXAMINATION FOR COLLEGE FRESHMEN AND HIGH SCHOOL SENIORS

The Thurstone examination[1] was also first published in 1919. It appears in yearly editions. It is No. IV of a series of six tests compiled by Thurstone while in the Carnegie Institute of Technology in Pittsburgh. The test is of the omnibus type; that is, various kinds of tasks are mixed together and after the test is started it is not stopped until the end of the test. The Otis Self-Administering Test is an example of this type of test. Thurstone has arranged the elements on a cyclical basis. After one group of questions of the information, analogies, completion, reasoning, proverbs, and number series completion type, the same cycle is repeated. The test consists of 168 questions or elements to be done in a time limit of thirty minutes.

TABLE 23

THURSTONE PSYCHOLOGICAL EXAMINATION FOR HIGH SCHOOL GRADUATES AND HIGH SCHOOL SENIORS — NO. IV: NORMS

SCHOOL	SCORE
Freshmen in Engineering Schools	89
Freshmen in Liberal Arts Colleges	87
Students in Normal Schools	75

SCHOLASTIC APTITUDE TEST

The College Entrance Examination Board prepared in 1926 under the direction of C. C. Brigham an intelligence test which they call the "Scholastic Aptitude Test." The test is very closely

[1] Published by C. H. Stoelting Company.

guarded and no copies have been circulated for inspection. However, a practice form which is distributed to each candidate may be used to give an idea of the nature and difficulty of the test.

TABLE 24

SCHOLASTIC APTITUDE TEST: TESTS, TIME ALLOWANCE, AND NUMBER OF ITEMS

TEST	TIME	NO. ITEMS
1. Definitions	9	30
2. Arithmetical Problems	8	20
3. Classification	6	40
4. Artificial Language	9	37
5. Antonyms	10	48
6. Number Series Completion	9	25
7. Analogies	6	40
8. Logical Inference	10	40
9. Paragraph Reading	30	50
	97	320

The preliminary results show that the test is highly accurate. A novel feature of the first test is the inclusion of supplementary experimental material at the end of the test which may be studied and from which selections may be made in making up tests in succeeding years.

RULES FOR GIVING A TEST

1. Before giving a test one should make sure that he knows how to give it. The best way to learn how to give a test is to take the test. Before any new test is given the examiner should read over the directions and if possible try them out on the family or some intimate friend.

2. Tests may be given at any time of the day or week. Although perhaps desirable, it is not at all necessary to give tests at the beginning of the day. They may be given late in the school

session with the confidence that the same results will be obtained as if they were given earlier.

3. Tests may be given in as large groups as seating facilities will accommodate. A group of 100, 200, or 500 may be as satisfactorily tested as a small class of 15 or 20, if there are sufficient assistants.

4. In seating pupils see that there is a space between seats or desks to lessen the tendency to cheat. In no case should pupils sit in double seats, or side by side. The space of an aisle should at least separate adjacent pupils. Desks or seats with arms are desirable, but if the tests are given in an assembly hall, lap boards may be used.

5. Have the pupils supplied with soft pencils—not with pens. Preferably each child should have two sharp pencils. If this is not possible, the examiner should have at hand an extra supply of sharpened pencils. Have all desks cleared of books and papers.

6. In giving tests to a group of 50 or more it is well for the examiner to have one or more assistants to help preserve order and to attend to the distributing of materials.

7. A good plan for getting the papers distributed is to place enough papers on the front desk of each row and then let the class pass the papers back, each one keeping one paper as he passes them along.

8. Do not excite the pupils by disorderly methods of seating or distributing materials, or by referring to the test as a "brain test" or an "intelligence test," or by impressing them with the importance of the test. Better say nothing at all about the test or its purpose — the pupils will discover that as they take it.

9. Read the directions slowly and distinctly with proper expression and emphasis. Otis suggests pausing two seconds between each sentence in giving directions.

10. Follow the directions verbatim. Do not interpolate your own remarks.

11. Let your voice be natural. Do not shout, yet at the same time speak so that every one can hear distinctly. The technique of group testing was developed in the army and naturally in the army more or less rigid military procedure was followed with commands such as "Attention!" snapped out in military style. This military manner of giving group tests has more or less persisted. In some ways it is a good thing. It insures snap and precision in the starting and stopping of tests and brings the class up to a certain level of obedience and attention. Yet the military manner of giving tests may be overdone with undue shouting and harshness.

12. Once the test is started do not permit questions. Be on the watch, however, for a pupil in distress, as for instance with a broken pencil, or an imperfectly printed test paper. Such accidents should be speedily remedied during the examination.

13. Have only one test exposed at a time by requiring the pupils to fold their test books back.

14. During the course of the test, make sure that there are no interruptions. A good plan is to have a sign on the outside of your door saying "Examination inside. Please do not disturb." One should make sure that during the examination another class will not have recess and laugh and shout beneath the windows. Give no help to the pupils and see that they get no help from one another.

15. A person giving a test which must be accurately timed should by all means use a watch with a second hand. It is impossible to time a test accurately with a watch which has no second hand. In starting a test it is not necessary to wait until the second hand comes to sixty. First get the class prepared by saying "Ready!" and then perhaps three seconds later say, "Go!" which is the signal for the class to start working. At the instant that you say "Go!" note carefully the second hand and write on a slip of paper figures such as 9:21, which indicate that the signal was given at 9 minutes and 21 seconds past the hour by your watch.

If the test is to run for two minutes, you should add 2 to 9 and write 11 : 21 on the slip of paper, indicating that you are going to stop your class at exactly 11 minutes and 21 seconds past the hour. Be careful to stop them promptly at the precise instant. One cannot stress too strongly the necessity for obtaining accurate timing in the giving of standardized tests. One should consider himself a scientist, in giving a standardized test, and should hold himself up to scientific standards of accuracy.

Directions for Scoring a Test

Each test has its detailed directions for scoring. A few cautions may be mentioned here.

1. Have the scoring done by teachers or clerks if you do not wish the pupils to become too familiar with the tests — that is, if you expect that they will ever be tested again with the same test, even another form.

2. Score the same test in all the papers before passing to the next test. This is a much more economical procedure than scoring each paper completely as you come to it.

3. Systematize the procedure. It is probably better to have the tests corrected at a stated time and place than to allow teachers to take them home. At any rate, a definite schedule when the corrected tests are due should be posted, otherwise the correcting will drag indefinitely.

4. Save the eyes as much as possible by bringing the key as near as one can to the answers. Keep the key on the same level on the page as the answers. Use stencils when provided. A stencil is a scoring key printed on transparent paper, or a key made by cutting holes through a piece of cardboard that will enable the scorer to see whether the correct element has been underlined. One must take care not to give credit when more than one element has been underlined.

5. Score the rights with a plus and the wrongs with a check.

Sometimes it is more economical to mark simply the wrongs, count them, and subtract from the total number attempted in order to get the number of rights. Judgment must be used as to the best method of scoring any test.

6. Give no credit for omitted answers.

7. In a test scored by the formula right minus wrong give no score less than zero.

8. Ascertain the intent of the examinee in doubtful cases. For example, if a pupil evidently has corrected his own error, give him credit for this.

9. Give no credit if two or more answers are given to a certain problem. Count as omitted if both true and false are underlined in a true-false test.

10. It is safest to have the additions of the sub-tests checked.

These are general rules. *The special rules governing the scoring of any test should be read before starting to score a set of papers.*

REFERENCES

Breed, F. S. — "The Status of Intelligence Tests," *School Review*, 30 : 242–244, April, 1922.

Cobb, M. V. — "The Limits Set to Educational Achievement by Limited Intelligence," *Journal of Educational Psychology*, 13 : 449–464, 546–555, Nov., Dec., 1922.

Colvin, S. S. — "The Present Status of Mental Testing," *Educational Review*, 64 : 196–206, Oct., 1922.

— "The Use of Intelligence Tests," *Educational Review*, 62 : 134–148, Sept., 1921.

Coxe, W. W. — "Norms for the Otis Group Intelligence Scale," *Journal of Educational Research*, 3 : 313–314, April, 1921.

Dickson, V. E. — *Mental Tests and the Classroom Teacher*, ch. 3. World Book Company, Yonkers; 231 pp., 1923.

Freeman, F. — *Mental Tests.* Houghton Mifflin Company, Boston, 1926.

Gregory, C. A. — *Fundamentals of Educational Measurement*, chs. 3–4. *Ref.* Ch. 2.

HINES, H. C. — *A Guide to Educational Measurements*, chs. 7–13. Houghton Mifflin Company, Boston; 270 pp., 1923.

JORDAN, A. M. — "The Validation of Intelligence Tests," *Journal of Educational Psychology*, 14 : 348–366, 414–425, Sept.–Oct., 1923.

LINCOLN, E. A. — *Beginnings in Educational Measurement*, ch. 5. J. B. Lippincott Company, Philadelphia; 151 pp., 1924.

MILLER, W. S. — "The Administrative Use of Intelligence Tests in the High School," TWENTY-FIRST YEARBOOK OF NATIONAL SOCIETY FOR THE STUDY OF EDUCATION, PART 2. Public School Publishing Company, Bloomington, Illinois; pp. 189–222, 1922.

MONROE, W. S., DEVOSS, J. C., and KELLY, F. J. — *Educational Tests and Measurements*, ch. 9. *Ref.* Ch. 2.

OTIS, A. S. — "An Absolute Point Scale for the Group Measure of Intelligence," *Journal of Educational Psychology*, 9 : 239–261, 333–348, May–June, 1918.

PINTNER, R. — *Intelligence Testing*. Henry Holt & Company, New York: 406 pp., 1923.

PRESSEY, S. L. and L. C., *Introduction to the Use of Standard Tests*, chs. 11–12. World Book Company, Yonkers; 263 pp., 1922.

RAUTH, J. W. — "An Evaluation of Some Group Intelligence Tests," *Catholic Educational Review*, pp. 209–214, April, 1924.

THORNDIKE, E. L. — "The Improvement of Mental Measurements," *Journal of Educational Research*, 11 : 1–12, Jan., 1925.

— "Intelligence Examinations for College Entrance," *Journal of Educational Research*, 1 : 329–337, May, 1920.

— *The Measurement of Intelligence;* Teachers College Bureau of Publications, Teachers College, Columbia University, New York, 1926.

THORNDIKE, E. L., and Others — "Intelligence and Its Measurement," *Journal of Educational Psychology*, 12 : 123–147, 195–216, 271–275, March, April, May, 1921.

THURSTONE, L. L. — "Cycle-Omnibus Intelligence Test for College Students," *Journal of Educational Research*, 4 : 265–278, Nov., 1921.

— "Mental Tests for College Entrance," *Journal of Educational Psychology*, 10 : 129–143, March, 1919.

TRABUE, M. R. — *Measuring Results in Education*, ch. 18. American Book Company, New York; 492 pp., 1924.

TRABUE, M. R., and STOCKBRIDGE, F. — *Measure Your Mind*, Doubleday, Page & Company, Garden City, New York; 349 pp., 1920.

WHIPPLE, G. M. — "Intelligence Tests and Their Uses," TWENTY-FIRST

YEARBOOK OF NATIONAL SOCIETY FOR THE STUDY OF EDUCATION, Public School Publishing Company, Bloomington, Illinois; 288 pp., 1922.

WYLIE, A. T. — "A Brief History of Mental Tests," *Teachers College Record*, 23 : 19–33, Jan., 1922.

YERKES, R. M. — "*Psychological Examining in the United States Army: History and Organization of the Psychological Examinations and the Materials Used in the United States Army*," MEMOIRS OF THE NATIONAL ACADEMY OF SCIENCES, VOL. 15; Government Printing Office, Washington, 1921.

YOAKUM, C. S., and YERKES, R. M. — *Army Mental Tests*, Henry Holt & Company, New York; 303 pp., 1920.

YOUNG, K. — "The History of Mental Testing," *Pedagogical Seminary*, 31 : 1–48, March, 1924.

CHAPTER V

STANDARDIZED TESTS FOR THE TEACHER OF ENGLISH

In a book of this type, which proposes to introduce the concepts of educational measurements to the high school teacher, to suggest devices and procedures, and to show some of the bearings of measurements on teaching and administrative problems, there is a limit to the detail into which one should go in the discussion of available standardized tests. Accordingly typical tests will be described in each subject without making any pretense at being complete or exhaustive. Mention of many valuable tests will be omitted, and only those tests will be described which seem most commendable. In different sections of the country there are fashions in tests, usually taken from some nearby school of education. There are criteria for judging tests, but they are difficult of application, so that in many cases personal preference is the only guide. If my readers prefer some other test, let them compare their favorites with the ones described in this book according to criteria to be given in a subsequent chapter and decide on their respective merits.

Before describing the tests in English let us pause for a moment to ask what are the objectives of the course in English in high school.

Objectives of Instruction in English in High School

1. Ability to communicate in speech.
2. Ability to communicate in writing.

All expression in writing demands correctness as to formal details, namely, a legible and firm handwriting, correct spelling, correctness in grammar and idiom, and observance of the ordinary

rules for capitalization and punctuation. The writer should make an effort to gain an enlarged vocabulary, a concise and vigorous style, and firmness and flexibility in constructing sentences and paragraphs.

3. Knowledge of books by better authors.
4. Power to read thoughtfully and with appreciation.

These objectives of the course in English have been stated very briefly. Ability in speech does not lend itself readily to standardized testing. Standardized tests are pencil and paper tests. It seems almost necessary to have them so in order to obtain the necessary objectivity. In order satisfactorily to measure oral composition a list of samples scaled in order of excellence would be necessary, but one cannot solidify speech to make this possible. Teachers who attempt to rate or mark oral compositions carry in their minds standards of excellence by which to make comparisons, but we know that such a method gives very unsatisfactory results. There seems to be no way of measuring scientifically this important objective in English.

Measurement of Composition Ability

For the measurement of ability to communicate in writing there are many tests and scales which have proved serviceable. The heart of this problem, the general ability to write a composition, was one of the first to be attacked in the whole movement in educational measurements. In 1912 Milo B. Hillegas published his "Scale for the Measurement of Quality in English Composition for Young People." This was one of the first of the quality scales in education. It consists of ten compositions selected from 7,000 samples graded and arranged in order of merit by the combined judgment of 202 qualified judges. Each degree of merit is represented by one composition. The compositions vary in length and in type and are on different subjects. Two are artificial compositions. One of these is of zero merit, a sample of writing

of "just not any" value. Other samples were written by high school pupils and college freshmen. Although the original scale lacked usefulness because of the nature of its selections, it was very carefully constructed and has been the basis of most of the best of the more recent scales. There have been several derivations from this scale, including the "Nassau County Supplement to the Hillegas Scale" and the original "Hudelson English Composition Ability Scale."

HUDELSON TYPICAL COMPOSITION ABILITY SCALE

The composition scale which seems to be most generally useful to-day is the "Hudelson Typical Composition Ability Scale." Its author is Earl Hudelson. This scale was first published in 1923. It may be used throughout the whole school career, from first grade through the university. Hudelson went to great pains to secure a topic which gives the most typical results in school work. All the samples of the scale are on this single topic. Some topics such as "How I Learned a Lesson" yield better than average compositions; many topics are difficult for children to write on. The topic chosen for the Typical Scale was "A Snowball Fight on Slatter's Hill," a story taken from T. B. Aldrich's *Story of a Bad Boy*. The use of the scale can best be described by repeating the directions for giving a composition test just as they are given by Hudelson.

When your pupils have been supplied with writing materials, say to them, "I am going to read you a story. Please listen carefully while I read it, for when I am through I want you to write it in your own words. I have already written on the blackboard the title of this story and the proper names that occur in it." Then read clearly the following story:

A SNOWBALL FIGHT ON SLATTER'S HILL

Slatter's Hill, or No Man's Land, as it was generally called, was a rise of ground covering perhaps an acre and a quarter, situated on an imaginary line marking the boundary between the North End and the South End of town.

One evening twenty or thirty of the North Enders quietly took possession of Slatter's Hill, and threw up a strong line of breastworks. Fancy the rage of the South Enders next day, when they spied our snowy fort, with Jack Harris's silk handkerchief floating defiantly from the flagstaff! In less than half an hour it was known all over town, in military circles at least, that the South Enders intended to attack the fort that Saturday afternoon.

At two o'clock all the fighting boys of the North End, and as many recruits as we could muster, lay behind the walls of Fort Slatter, with three hundred snowballs piled up, awaiting the approach of the enemy. The enemy was not slow in making his approach, — fifty strong, headed by Mat Ames. Our forces were under the command of General J. Harris.

The thrilling moment had now arrived. The fort opened fire first, — a single ball from the hand of General Harris taking General Ames in the very pit of his stomach. A cheer went up from Fort Slatter. In an instant the air was thick with flying snowballs, in the midst of which we dimly saw the storming parties sweeping up the hill, shoulder to shoulder.

Not more than a dozen of the enemy succeeded in reaching the crest of the hill; five of these clambered upon the icy walls, where they were instantly grabbed by the legs, jerked into the fort, and made prisoners. Twice we were within an ace of being driven from our stronghold, when General Harris and his staff leaned recklessly upon the ramparts and hurled the besiegers heels over head down the hill. The rest retired confused and blinded by our well-directed fire.

At sunset the garrison of Fort Slatter was still unconquered, and the South Enders marched away whistling "Yankee Doodle," while we cheered and jeered them till they were out of hearing. When General Harris, with his right eye bunged up, said, "Soldiers, I'm proud of you!" my heart swelled in my bosom.

The battle passed into a legend; and afterwards, when later instances of pluck and endurance were spoken of, the boys would say, "By golly, you ought to have been at the fight on Slatter's Hill!"

Collect all papers promptly at the end of fifteen minutes. Before reading this to your pupils the following words besides the title of the story should be written on the blackboard:

Slatter's Hill
North End
South End
Jack Harris's
Mat Ames

When you are ready to score the themes, place the scale before you, review it, read the first theme to be scored, compare it with the scale samples one by one, and give it the value of that scale sample to which it most nearly compares in general composition merit. At first you will find that you are able to give only the values found in the scale itself, for instance, 6.0 or 7.0. After having had experience in grading themes by means of the scale, you will find that you can give values between the values on the scale; for instance, the theme may be better than 6.0 and not so good as 7.0, in which case you would give it a value like 6.2, 6.5, or some intermediate value, as your judgment may dictate. When you have become expert with the use of the scale, you will not need to begin with the lowest sample and consult each step in order, but can go at once to that part of the scale which contains samples approximately equal in quality to the theme to be scored, and by studying the two or three consecutive samples in that level of the scale you can quickly assign the theme its appropriate score. Grade norms are as follows:

TABLE 25

HUDELSON TYPICAL COMPOSITION ABILITY SCALE: GRADE NORMS (JANUARY)

Grade	Quality
IV	3.0
V	3.6
VI	4.2
VII	4.7
VIII	5.3
IX	5.5
X	5.9
XI	6.3
XII	6.7

These norms are well established, being the result of many years of experience with the Hillegas Scale in school surveys and the like.

In order to provide controlled practice for use of the scale, Hudelson provides practice exercises in the manual of directions. These should be rated by the teacher and then her values compared with the key given on page 16 in the manual.

One should inform himself from the exercises whether he is scoring too severely or too leniently and how much on the average he is in error. These facts may be obtained objectively. The following illustrates how this may be done.

Your problem is to find your average error in scoring and whether you rate too severely or too leniently and how much.

TABLE 26

SHOWING COMPUTATION FOR DETERMINING AVERAGE ERROR AND DIRECTION OF ERROR IN RATING A COMPOSITION BY MEANS OF A COMPOSITION SCALE

THEME	SCORE	YOUR SCORE (HYPOTHETICAL)	DIFFERENCES WITHOUT SIGNS	POSITIVE DIFFERENCES	NEGATIVE DIFFERENCES
D1	6.5	6.5	.0		
D2	2.6	3.0	.4	+ .4	
D3	7.9	7.0	.9		− .9
D4	3.2	4.5	1.3	+ 1.3	
D5	0.8	2.0	1.2	+ 1.2	
D6	4.5	4.0	.5		− .5
D7	1.25	1.7	.45	+ .45	
D8	1.1	1.5	.4	+ .4	
D9	8.0	6.5	1.5		− 1.5
D10	2.2	2.8	.6	+ .6	
			10)7.25(.725	+ 4.35	− 2.9

+ 4.35
− 2.9
10)1.45(.145

Average error in scoring = .725
Scoring too leniently by
an average amount of .145

Steps: 1. Place your own scores opposite the true scores as in the above example.
2. Take the difference without regard to signs and place in the next column.
3. Find the sum of this column.
4. Divide this sum by 10 (the number of themes graded). This quotient is your average error in grading these ten compositions.
5. If your score is greater than the true score, subtract again and place the difference in the column headed "Positive Differences." Place a plus sign before each of these differences. If your score is less than the true score, subtract and place the difference in the column headed "Negative Differences." Place a minus sign before each of these differences.
6. Add each of these columns. Call the sum of the Positive Differences column plus, and the sum of the Negative Differences column minus.
7. Subtract the sums of these two columns. Give this difference the sign of the larger number.
8. Divide by 10, the number of themes graded. If this result is plus, you have graded too leniently; if this result is minus you have graded too severely.

NOTE: If your score is the same as the true score, the difference is 0 and will not belong in either the positive or the negative column. One divides by the total number of compositions scored in any case.

One should not consider himself competent in grading a composition until he is able to assign values which are not more than .5 from the true value.

In grading a pupil's theme the teacher should not rely on a single rating of one composition. If possible, at least three independent ratings should be obtained from different persons before one can be sure of an individual's "score" or composition ability. Also the rating should not depend on a single composition, but two or more. If these requirements seem to put undue burden on the classroom teacher, it only shows that it is not a simple

matter to gauge accurately composition ability.[1] Even so, the use of the scale will make the judgment of a composition more accurate than a judgment with no objective guide to follow. In group testing, however, one need not be so painstaking. Compositions may be assigned values as found on the scale without the interpolation of intermediate values and the average of these values will be a fairly accurate index of class ability.

If one consults the norms given on page 88, he will see that yearly progress or gain in the high school is only about 0.4 of a point on the scale. It must be evident, then, that pupil progress within a year is extremely difficult to measure unless many composition samples are rated very accurately. The same is not true of class progress, for the mean of a group is more stable than the scores of individuals. If the scale is to be used oftener than once a year, it is well to vary the assignment. Hudelson gives ten supplementary assignments in the teacher's handbook accompanying the scale. It must be evident from this that the scale is not adapted to daily, weekly, or even monthly use.

There is a value of quite a different nature, however, that may come from the use of the scale. Ability to judge a composition's merit may help to increase ability to write a composition; and vice versa, ability to write a better composition may help to increase one's ability to judge composition value. Chou[2] found that there is a high relationship between ability to write and ability to judge (correlation .71) and that there is strong presumption that training in one results in increased ability in the other — both are valuable abilities. Accordingly the scale should be put into the hands of the pupils and they should be given practice in judging their own and others' compositions.

[1] Chou says: "Ten compositions written by each pupil and scored by three trained judges on a standardized scale (or thirty judgments given to each pupil's work) was deemed sufficient to represent the true composition ability of the pupil examined." *The Measurement of Composition Ability*, p. 62, 1922.

[2] *Op. cit.*, pp. 92, 93.

Even if the cause-effect relation between judging and writing is slight, such practice makes the child appreciate better the teacher's appraisal of him and makes him realize that he is striving not to win the applause of the teacher, but is aiming toward a very objective goal.

LEWIS COMPOSITION SCALES

For measurement of special types of composition commonly considered important Erwin E. Lewis has prepared scales,[1] and these should be used for those special types of composition. The scales measure quality of order letters, letters of application, narrative social letters, expository social letters, and simple narration. This is a derivative of the Hillegas scale and Hillegas norms may be used.

VAN WAGENEN COMPOSITION SCALES

The teacher who wishes to measure various phases of composition may be interested in the "Van Wagenen Composition Scales." There are three separate scales for the measurement of exposition, narration, and description. For each composition there is provision for rating separately Thought Content, Structure, and Mechanics, three distinct and vital phases of writing. These scales are also Hillegas derivatives and were published in 1923.[1] Van Wagenen's description[2] of how to rate compositions is as follows:

> Compare the composition to be measured with the specimens on the scale from the point of view of Thought Content. Assign to the composition the Thought Content value of that specimen which the composition most closely resembles; or in case it is better than one specimen and not so good as the next high one, assign to the composition an intermediate value. Then rate the composition for structure and then for mechanical errors in the same manner.

[1] Published by The World Book Company in 1923.

[2] Cf. *Manual of Directions*.

In rating for Thought Content in Exposition, take into consideration:

Adherence to subject;
Interest in treatment;
Continuity of thought;
Clearness of perception;
Discrimination in selection of words.

In rating for Thought Content in Narration, take into consideration:

Sufficient explanation of the situation;
Naturalness and appropriateness of dialogue (if used);
Clear progress of narrative to a definite conclusion;
Use of suspense or surprise;
Descriptive touches;
Adequacy and variety of diction.

In rating Thought Content in Description, take into consideration:

Maintenance of point of view (both physical and mental);
Vividness of picture;
Emotional reaction;
Vigor and originality of diction.

In rating for sentence and paragraph structure, take into consideration:

Unity;
Coherence;
Emphasis;
Variety and complexity of sentences.

The writer who uses many complex sentences shows a greater maturity of mind than the one who uses very simple or unnecessarily compounded sentences, even though, from the very fact of the greater complexity, he may make more actual mistakes in structure.

In rating for mechanical errors, take into consideration:

Spelling;
Punctuation (only in cases of actual error, not cases where punctuation is optional);
Capital letters;
Grammar (only cases of actual error, not matters of preference);
Paragraphing (only cases of actual error, not matters of preference).

Here, too, one must take into consideration the range of vocabulary and complexity of expression. For instance, it is a more fundamental error to misspell "receive" than to misspell "psychological." Punctuation of very simple sentences would give less opportunity for error than that of complicated sentences or conversation.

Van Wagenen suggests getting a rating for general merit by use of a formula. If *GM* stands for the rating in General Merit, *T* stands for the rating in Thought Content, *S* stands for the rating in Structure, and *M* stands for the rating in Mechanics.

$$GM = \frac{4\,T + 2\,S + 1\,M}{7}.$$

Hudelson's[1] criticism of these scales is worth quoting: "While these scales represent a worthy attempt to analyze composition writing for diagnostic purposes, they render judgments confusing and difficult, if, as is customary with teachers, the separate evaluations are combined into one general score." He also adds, "compositions are less reliably judged when analyzed merely into thought content, structural and mechanical aspects than when they are considered from the standpoint of general merit."

The teacher who wishes to measure the progress of composition ability more frequently than once a year should have recourse to the tests of handwriting, spelling, correct usage, punctuation and capitalization, vocabulary, and ability with sentences. For the larger aspects of style and the structure of larger forms, standardized tests have not yet been constructed.

Composition scales are not considered with the same favor today as they were several years ago. In the first place a single teacher's estimate of a set of compositions is too unreliable to be usable. In the second place the use of composition scales tends to stifle individuality, causing a teacher to overlook originality of thought and expression and to stress conformity, conventionality, and mechanical correctness.

[1] *Op. cit.*, p. 52.

Spelling Tests

For the testing of spelling in high school sixteen spelling scales[1] have been constructed by Earl Hudelson, F. L. Stetson, and Ella Woodyard under the direction of Thomas H. Briggs and T. L. Kelley. There have been excellent scales made for use in the elementary school, notably the "Ayres Spelling Scale" and the "Iowa Spelling Scale." Ayres has given us the 1,000 most common words for the spelling vocabulary. The workers mentioned above determined the second and third thousand most common words for the spelling vocabulary, partly on the basis of previous word counts and partly on additional counts. These 2,000 words were put into 23 lists, 90 words in a list, with nine anchor words common to all 23 lists, and in addition one word found in both the Ayres and Buckingham scales, making 100 words in all. These were given to pupils in secondary schools from grades seven to twelve, each list being spelled by from 160 to over 1,200 pupils. From this testing the difficulty value of each word was found. Using these difficulty values as a basis, sixteen spelling scales of 20 words each were constructed, the words being incorporated into sentences. Further testing was carried out to make certain that the scales were equal in difficulty. The first twelve scales represent one level of difficulty so that each word in each of the twelve lists is equal in difficulty to the corresponding word in each of the other eleven lists. The words in lists 13 to 16 are more difficult, but there is an overlapping in difficulty between the two groups. The corresponding words in each of these last four scales are of equal difficulty. The norms for the test are given in the table on page 96. That is, in the seventh grade an average pupil will be able to get 65.9% of the words on list I correct (or any of the first twelve lists). More elaborate norms accompany the scales so that a teacher may construct her own test and determine the norms for it.

[1] Published by Bureau of Publications, Teachers College, Columbia University.

TABLE 27

SIXTEEN SPELLING SCALES: FEBRUARY NORMS GIVEN IN TERMS OF PER CENT

GRADE	LISTS I–XII	LISTS XIII–XVI
7th	65.9	34.8
8th	73.8	45.0
9th	80.0	53.9
10th	85.1	61.5
11th	88.7	67.1
12th	91.2	72.1

Spelling is a difficult phase of English in high school because the work lacks motivation for the students. The student does not see where he is going nor does he see any gain or any advantage in his study. As spelling is usually treated in high school the pupil is presented with endless lists of difficult and often unfamiliar words. The work may be given some reality by making it grow out of actual errors in composition writing, but standardized testing of spelling will give the work real motive. The accuracy of the scales is such that increased spelling skill may be readily detected and pupils may be given assignments in terms of actual goals to be reached in future spelling tests.

LANGUAGE USAGE, PUNCTUATION, CAPITALIZATION, AND ABILITY WITH SENTENCES

THE PRESSEY DIAGNOSTIC TESTS IN ENGLISH COMPOSITION

These tests have been selected, out of the many available, as representative of tests designed to give a diagnostic measure of ability in English composition.[1] There are four separate tests in this series. Each test comes in two forms.

[1] These tests are published on separate sheets by the Public School Publishing Company (1923).

NAME	AUTHORS
1. Capitalization	S. L. Pressey and Mrs. E. V. Bowers
2. Punctuation	S. L. Pressey and Helen Ruhlen
3. Grammar	F. R. Conkling and S. L. Pressey
4. Sentence Structure	F. R. Conkling and S. L. Pressey

Test of Capitalization

This test consists of twenty-eight sentences and each sentence lacks capitalization aside from the capital at the beginning of the sentence. The pupil is instructed to draw a line under each letter that should be a capital. Five minutes is allowed.

The test is arranged to give the teacher detailed diagnostic or analytic information with regard to each pupil's knowledge of good usage in capitalization. It is based on an elaborate statistical study regarding the frequency with which capitals are used for various purposes in periodicals, newspapers, and business letters. For convenience in analyzing the results, the sentences have been grouped to exemplify certain rules, which formulate important current practices as revealed in this study. The rules are given below in order as illustrated in the test. In the test, each rule is illustrated by four sentences.

1. Capitalize the names of persons, with their titles; however, do not capitalize titles when they are not part of a name.
2. Capitalize the names of countries, states, cities, streets, buildings, mountains, rivers, oceans, or any word designating a particular location or part of the world; however, do not capitalize the points of the compass.
3. Capitalize the names of business firms, schools, societies, or other organizations.
4. Capitalize words derived from the names of countries, and places, organizations or persons.
5. Capitalize the first word and all other important words in titles (and sub-titles and headings) of themes, magazine articles, poems, books, of laws or governmental documents, and the trade names of commercial products.
6. Capitalize the days of the week, the months of the year, and holidays and church days; however, do not capitalize the seasons.
7. Capitalize the first word of every sentence, the first word of every line

of poetry, and the first word of a direct quotation; however, if the quotation begins in the middle of a sentence (or is indirect) do not use the capital.

The test investigates, systematically, knowledge of these rules, with the exception to the rule that a sentence must begin with a capital; knowledge of this rule has been taken for granted, but it has, nevertheless, received emphasis by special mention in the directions. Also, no mention is made, in rules or test, of "I" or "O." "O" is omitted as too infrequent, in the written work of the average individual, to deserve inclusion, "I" as too generally known by children in the grades covered by the test, to need mention. The various cases where capitals are used to show respect (as in mention of the Bible, of the Deity, of high government officials of governmental bodies, or reference to one's school or college, business house, or some other organization for which one wishes to show special regard) have also been omitted, as infrequent in ordinary written work, or matters of variant practice.

Test of Punctuation

This test consists of thirty sentences, all of which lack certain punctuation marks. The pupil is to supply the missing punctuation, changing periods to exclamation points or interrogation marks where necessary. Ten minutes are allowed.

The test is based upon a detailed analytical study of current usages in punctuation as shown in business letters, newspapers, and magazines. It covers systematically common practices as thus determined. The following rules are offered as a formulation of usage as revealed by this study.

Full stops: Use a question mark after a direct question (not after an indirect question). Use an exclamation point after a sentence, exclamation, or interjection to show strong emotion or surprise. At the end of all other sentences use a period. Use a period also after abbreviations and initials.

Pauses within a sentence: Use a comma (1) to set off slightly parenthetical, or inserted, words, phrases, or clauses, (2) to set off clearly introductory words, phrases, or clauses at the beginning of a sentence, or obviously added elements at the end, (3) to separate words or phrases in a series, (4) to separate clauses joined by "and," "but," "for," "as," or any other simple conjunction. The comma is also used (5) to separate the parts of a date or an address, (6) to introduce a short quotation, and (7) after the complimentary close of a letter.

Use a semicolon (1) between clauses of a compound sentence that are not joined by a conjunction, and (2) to make prominent a division within a sentence when the parts separated are very long, or have commas within themselves.

Use a colon (1) after words, phrases, or sentences serving as a formal introduction to something that follows, as a list or a long quotation. Use the colon also (2) after the salutation in a business letter and (3) between the hours and minutes in a statement of time.

Use a dash to indicate a marked break in the progress of thought in a sentence, as when an explanatory element is obviously inserted. The parenthesis may also be used for this last purpose.

Special marks: Use the apostrophe (1) to indicate the omission of a letter or letters in a word, and (2) to indicate the possessive case. However, possessive personal pronouns (its, his, theirs, yours, ours) do not take the apostrophe.

Use quotation marks (1) to inclose a direct quotation (not an indirect quotation) and (2) to indicate the title of a theme, a short story, a magazine article, a poem, or play. However, italicize the names of books and magazines (italics to be indicated by underlining). Quotation marks may also be used (3) to call attention to technical, foreign, or unusual words, or words used with some special — as ironical or humorous — meaning.

The test investigates systematically knowledge of these rules. In each group of six the first two sentences have to do with the use of commas; the third, with use of semicolon, colon, dash, or parenthesis; the fourth, with full stops; the fifth, with the apostrophe; and the sixth, with the use of quotation marks.

The test does not include consideration of the salutation and complimentary close of letters.

Test of Sentence Structure

The test consists of statements in groups of four. In each group one statement is not well expressed — has in it something which does not make a good sentence. The pupil is to mark the line which has the poor sentence by a cross. Example:

☐ He finished school last June. Since that time he has done nothing.
☐ We spent the afternoon in the park: it is very pleasant there.

☒ John had a position in a store. Which he left.
☐ He found a dollar bill which some one had dropped in the aisle.

There are twenty-four groups of statements.
Twelve minutes are allowed.

The test is arranged to give the teacher detailed diagnostic or analytic information regarding each pupil's ability to recognize the most common mistakes in sentence structure. It is based on several studies regarding the frequency of such errors in the written work of both children and adults. For convenience in analyzing results the items of the test have been grouped according as they exemplify certain simple rules for the avoidance of the common errors. The rules are as follows:

1. Make certain, in the first place, that all sentences are real sentences; do not punctuate a part of a sentence — a phrase or a clause — as if this fragment were a whole sentence. (Items 1–4.)

2. Be careful not to put too much into one sentence. Be suspicious of long sentences with parts joined together by several *and's* or *so's*, or simply by commas. When such sentences are found, either (*a*) cut them up into a number of shorter sentences, (*b*) put into subordinate clauses or phrases any subordinate ideas, or (*c*) use semicolons in place of the commas, if the sentence permits of this handling. (Items 5–8.)

3. Be certain that all pronouns have a definite antecedent. (Items 9–12.)

4. Be sure that all modifiers have something to modify. (Items 13–16.)

5. Keep pronouns close to their antecedents, and modifiers close to the words they modify. (Items 17–20.)

6. Be careful not to use prepositions or pronouns when such use repeats a meaning already expressed. Be still more watchful, however, that such small words are not omitted when needed. After finishing any written work it is well to look over the paper to be certain that such omissions have not been made. (Items 21–24.)

Tests and rules cover all errors in this general field common in both junior and senior high school, so far as it is possible to determine these from the few studies regarding errors in sentence structure which have thus far been made. These studies agree in showing that the bulk of the errors, especially in the junior high school and first year of the senior high school, are of the few relatively simple types indicated above. It is further believed that of the more elaborate mistakes many will disappear if the rules given above — especially the second — are followed.

Test of Grammar

Sentences are given in groups of four. One of the four sentences in each group contains something which is not good English. The pupil is to indicate the wrong sentence by a cross. Example:

They were here.
We was going.
Where were you yesterday?
It was time to go.

There are thirty groups of sentences.
Fifteen minutes are allowed.

The test is arranged to give the teacher detailed diagnostic or analytic information regarding each pupil's ability to recognize the most common mistakes in grammar, using this term to cover primarily the subject of correct use of the few inflected forms remaining in the English language — errors in sentence structure are dealt with in a separate test. It is based on several studies regarding the frequency of grammatical errors in the written work of both children and adults. For convenience in analyzing results the items of the test have been grouped accordingly as they exemplify certain simple rules for the avoidance of the common errors. The rules are as follows:

1. Remember that a pronoun should agree in number with its antecedent, and a verb with its subject. Few mistakes will be made if the following points are kept in mind. (*a*) Two or more singular nouns joined by *and* require a plural verb and pronoun; if joined by *or* or *nor*, singular verbs and pronouns are to be used. (*b*) *There is* or *there are* should be used according as the subject that follows is singular or plural. (*c*) Intervening words in a different number must not be allowed to affect the agreement of pronouns or verbs. (*d*) *Each*, *every*, *either*, *somebody*, *any one*, *one*, and other similar pronouns are singular. (*e*) A singular noun used in a general way to refer to a class is also singular. (Items 1–10.)
2. Be careful about the tense forms of verbs. (*a*) Do not leave off the *ed* of past tense and participial forms of regular verbs; this mistake is most common with *ask*, *seem*, and *use*. (*b*) Be certain of the correct past tense and past participle of the following irregular

verbs: *come, see, run, do, go, give, bear, lead, grow, hear, ride, drink, burst.* (*c*) Remember that *drown, climb, drag,* and *attack* are regular. (Items 11–15.)

3. Certain forms are very commonly confused, usually because of similarity of sound. (*a*) Do not use *of* or *a* for *have* in such phrases as "might have been" or "ought to have gone." (*b*) Distinguish the following verbs: *lie* and *lay, sit* and *set, let* and *leave, can* and *may.* (*c*) Distinguish also the following: *there, their,* and *they're; there's* and *theirs; to, too,* and *two; a, an,* and *and.* (Items 16–23.)
4. Do not use adjectives for adverbs. (*a*) Do not use *good* for *well.* (*b*) Do not leave off the *ly* of adverbial forms. (Items 24–26.)
5. Few mistakes in the case of pronouns will be made if the following points are kept in mind. (*a*) Do not let words coming between a pronoun and the word governing it mislead as to the case. (*b*) Do not neglect the case of the pronoun in such phrases as *we boys.* (*c*) Always consider the implied construction after *than* and *as.* (Items 27–30.)

Tests and rules cover all common errors in this general field with the exception of mistakes in the number of nouns. The omission has been on the ground that a large proportion of these errors were not grammatical in nature — were mere slips of the pen (as in omitting the *s* of the plural), or mistakes in irregular plurals which were matters of spelling rather than grammar. Constructions well on the road to acceptance as idiomatic, such as "It is me," have not been included.

The mental operations on these tests are of the nature of proofreading. Quite obviously composition ability and proofreading ability are totally distinct. Willing[1] finds that the correlation between ability to write correct English and ability to make corrections in printed material is far from perfect. The tests are inexpensive and should be useful to the teacher who is specializing on one or another phase of composition. Norms are given for the high school grades, which represent the widest testing for obtaining norms of any educational test extant to-day. Over 130,000 cases representing every state in the union were averaged to provide norms for each of the four tests.

[1] Willing, M. H. — *Valid Diagnosis in High School Composition*, ch. 6.

TABLE 28

PRESSEY DIAGNOSTIC ENGLISH COMPOSITION TESTS: GRADE NORMS

	CAPITALIZATION						GRAMMAR					
GRADE	7	8	9	10	11	12	7	8	9	10	11	12
25 percentile	15.6	17.7	18.7	19.9	20.8	21.3	9.0	10.9	12.4	14.6	16.4	18.1
Median . .	18.8	20.6	21.5	22.8	23.3	23.8	12.0	14.6	16.5	19.0	21.2	22.7
75 percentile	19.2	23.2	24.0	25.3	25.4	25.7	15.6	18.5	21.2	24.9	23.4	25.0
	PUNCTUATION						SENTENCE STRUCTURE					
GRADE	7	8	9	10	11	12	7	8	9	10	11	12
25 percentile	7.3	9.2	10.2	12.9	14.7	16.0	9.0	10.9	11.1	14.0	15.4	16.6
Median . .	10.4	12.4	14.0	17.3	19.1	20.2	11.8	13.6	15.1	17.2	18.6	19.5
75 percentile	13.1	16.8	18.3	21.3	22.7	22.5	16.0	18.1	18.5	20.1	21.5	22.0

THE CROSS ENGLISH TEST

This test[1] was constructed by E. A. Cross. It comes in three duplicate forms *A*, *B*, and *C*, equalized for difficulty. This test was tried out and standardized in Cross' English classes in the Colorado State Teachers College. It consists of eight parts, as follows:

Part I. Spelling
II. Pronunciation
III. Recognizing a sentence
- *A*. Distinguishing between a complete and an incomplete sentence
- *B*. Judging whether material should be written as one or as two sentences
- *C*. Recognizing faulty sentences
- *D*. Judging whether a sentence as it stands should be broken into two sentences
- *E*. Recognizing faulty sentences

IV. Punctuation
V. Verb Forms
VI. Pronoun Forms
VII. Idiomatic Expression
VIII. Miscellaneous Faulty Expressions

[1] Published by World Book Company (1924).

Part I is a test of ability to recognize correct from incorrect spelling.

Directions: Place a check mark (√) in the parenthesis after each *correct* spelling, as in the sample

{	Inglish	( )
	English	(√)

From offhand observation Part I seems to be less valuable than some other parts of the test. No one knows how well proofreading ability (which this part tests) correlates with ability to spell. Certainly the correspondence is not perfect. Ability to spell well does not necessarily indicate ability to recognize readily incorrect spelling. Part II is intended to be a test of pronunciation. The directions are: "Place a check mark in the parenthesis nearest each correct pronunciation, as in the samples. Give careful attention to the position of the accent mark." Below this is a key to the diacritical marks used to indicate pronunciation. Following are samples of the exercise:

Samples	said	sād	( )	(√)	sĕd
	been	bĭn	(√)	( )	bēn
	does	do͞oz	( )	(√)	dŭz

This is probably more of a test of ability to read and comprehend the marks used and to discriminate between small differences than it is a test of ability to pronounce correctly. It is my surmise that the test measures only slightly what it is intended to measure, but the truth or untruth of this surmise rests on the finding of experimental evidence. The other parts of the test are of genuine value.

At the present time there are no norms for the high school grades, but the test may be used with profit in the high school. Norms should be given for each part separately. The manual of directions accompanying the test gives full directions for administering. The time allowance for the whole test is forty-five minutes. The

pupil proceeds through all the parts at his own rate, plenty of time being allowed in the forty-five minutes for him to do what he can. The key provided facilitates scoring. The manual of directions states that there is a practice effect of four points in a second test after a first test, which means that in comparing a second giving with norms an allowance should be made.

These tests may be used like a general composition scale to section pupils according to ability. They also have two other distinct uses. This test may be used to measure learning over short periods of time. For instance, if instruction has emphasized punctuation for a period of several weeks, the gain in ability to punctuate would be evidenced by the differences in score on tests given before and after training. Both the teacher and the pupil may thereby obtain an objective check on the learning. A second use is that of diagnosis. By a tabulation of the errors, stock may be taken of individual and class weaknesses which will give point to further emphasis in instruction. Indeed if a teacher were to use the test before instruction, then her teaching would be selective, giving emphasis to weaknesses. At present thoroughness is the aim of teaching whether equal emphasis is needed on all topics or not.

There are several other valuable tests which measure these same phases of language ability. Charter's "Diagnostic Language Tests" measure ability to recognize incorrect language usage and to correct it. These tests have no norms above the ninth grade, nevertheless they would probably be serviceable all through high school.

Briggs has constructed an "English Form Test" in two alternate forms, Alpha and Beta, which gives a diagnosis of ability in capitalization and punctuation. There are no norms for the high school on the test as a whole, but norms are given for grades seven, eight, and nine on each type of error in terms of per cent of error. Wilson has constructed a "Language Error Test" in the form of a story

containing many errors, mainly grammatical or idiomatic, which the pupil corrects. There are three different stories which have been carefully equated to give comparable scores.

Vocabulary and Reading Ability

HAGGERTY READING EXAMINATION, SIGMA 3

This test is selected as typical of reading tests for use in the high school. It has the added advantage of combining in a single booklet a vocabulary test, a test for comprehension of short sentence units and comprehension in larger paragraph units. The author is M. E. Haggerty. The test [1] has two forms, *A* and *B*, useful from grades five to twelve. The vocabulary test is of the conventional multiple choice type, the task being that of underlining the word meaning most nearly the same as the word to be defined. Test two, sentence reading, approximates the true-false test in type. It consists of a number of yes-no questions. The items in the test gradually become more complex in the structure of the question, in the vocabulary used, and in the information necessary to judge the answer correctly. Test three, paragraph reading, consists of a set of seven paragraphs with questions asked about the material in the paragraph. The paragraphs have been criticized as being passages that are from standard literature taught in high school, so that this may be more of a test of what has been learned than of ability to read unfamiliar material. This criticism probably is not justified. Alternative responses are printed to these questions and the pupil has to indicate the best or truest response out of the three or four presented. Method of scoring, time allowance, and maximum score are presented in the table following:

[1] Published by the World Book Company (1920).

TABLE 29

HAGGERTY READING EXAMINATION, SIGMA 3:
TESTS, METHOD OF SCORING AND POSSIBLE SCORE

TEST NO.	METHOD OF SCORING	MAXIMUM SCORE	TIME ALLOWANCE IN MINUTES
1	Rights	50	5
2	Rights − Wrongs	40	3
3	Rights × 2	54	20
		144	28

The norms given for the score on the whole test are:

TABLE 30

HAGGERTY READING EXAMINATION, SIGMA 3: GRADE NORMS

Grade . .	5	6	7	8	9	10	11	12
Score . .	31	50	68	76	84	90	96	102

The manual of directions for the Haggerty Test must be purchased separately.

It should not be necessary to persuade the teacher of the importance of vocabulary mastery and skill in reading in the work of the English class. Success, not only in English but in any high school subject, depends on these two important functions. So important are they that high ability in them is very indicative of high intelligence in general. To what extent vocabulary ability and ability to read are amenable to instruction in the high school is not known. However, they are stated to be important objectives of the work in English, and presumably skilful, well-directed teaching bears its results. Not only the English teacher but

every teacher can make no better single diagnosis of her pupil's ability than to measure his general vocabulary and reading ability.

There are other important vocabulary tests. The "Thorndike Test of Word Knowledge" is probably the best single English vocabulary test. Norms have been determined for grades 7, 8, and 9, but there are no norms for the upper high school years. This test comes in four alternative forms. The "Inglis Vocabulary Test," constructed by the late Professor Alexander Inglis, has a more difficult selection of words than the Thorndike Test, making it more generally useful in the later high school years. The Thorndike test contains 100 words, whereas the Inglis test has 150 words. This latter feature of itself should make the Inglis test the more reliable. The Inglis test is peculiar in that the items are placed in alphabetical order instead of in order of difficulty as in most educational tests. This feature makes no difference in the total score, but precludes any determination of the level at which the pupil begins to fail in his word knowledge.

There are also other reading tests adapted to the high school level of difficulty. The "Thorndike-McCall Reading Scales" are particularly valuable. These are derivatives of the original Thorndike Reading Scale and may be used from grades two to twelve. The norms begin to run pretty closely together in the last years of high school, making increase in ability less easy to measure with these scales, although one will find plenty of differences in reading ability on the scale in any high school class. This test comes in ten alternative forms, making it particularly useful if re-testing is planned. The Thorndike-McCall test measures comprehension alone. For high school pupils there is much waste material in the easy end of these scales. Some pupils tend to be careless on material that is obviously intended for those of less ability.

Van Wagenen has devised scales for measuring ability to read special high school content. There are scales for testing the ability to read thoughtfully history, general science, and English literature. It may be that reading ability is specialized in these upper years according to the type of material. "Monroe's Standardized Silent Reading Scale, Test III," enables one to get a measure of speed as well as comprehension in reading in the high school. The speed on the Monroe tests is not, however, a true speed of reading score as it includes the time for answering the questions as well as doing the reading. This time score should not be compared with the speed of reading obtained from reading consecutive material.

There are no tried standardized tests for testing appreciation in the field of English literature or for testing one's acquaintance with the content of standard literature that is commonly read in the high school. Abbott and Trabue[1] have devised an ingenious test which consists of poetry classics with three additional versions of each so like the original that close discrimination is necessary to tell the difference. The task is to determine which is the original.

These exercises measure the ability of college students to discriminate between different degrees of quality in English poetry. In the high school, where they are less reliable as tests, these exercises may be used as material for class work, especially for building up within students' minds a recognition of and desire for worthy rather than inferior poetic expressions. Series X and Series Y are each graduated, from Mother Goose to Milton or Browning, and are parallel to each other in difficulty at each step. Each series contains thirteen sets, each set consisting of four versions of the same selection from which the student is asked to select a "Best" and "Worst" version. Each series, when used as a test, requires approximately one class period for administration.

[1] Abbott, A., and Trabue, M. R. — "A Measure of Ability to Judge Poetry," *Teachers College Record* 22: 101–106, March, 1921.

REFERENCES

(Objectives in English)

Bobbitt, F. — *How to Make a Curriculum*, chs. 6, 16. Houghton Mifflin Company, Boston; 292 pp., 1924.

Hosic, J. F. — "Reorganization of English in Secondary Schools," U. S. Bureau of Education, Bulletin no. 2, 1917.

(Tests in Reading and Vocabulary)

Abbott, A., and Trabue, M. R. — "A Measure of Ability to Judge Poetry," *Teachers College Record*, 22: 101–126, March, 1921.

Brooks, F. D. — "The Reliability of Silent Reading Tests," *School and Society*, 19: 652, May 31, 1924.

Gates, A. I. — "An Experimental and Statistical Study of Reading and Reading Tests," *Journal of Educational Psychology*, 12: 303–314, 378–390, 445–464, Sept., Oct., Nov., 1921.

— "Study of Depth and Rate of Comprehension in Reading by Means of a Practice Experiment," *Journal of Educational Research*, 7: 37–49, Jan., 1923.

Gray, W. S. — *Deficiencies in Reading Ability: Their Diagnosis and Remedies*, ch. 16. D. C. Heath & Company, Boston; 420 pp., 1922.

Inglis, A. — "A Vocabulary Test for High School and College Students," *English Leaflet*, 23: 1–13, Oct., 1923.

Kelley, T. L. — "Thorndike's Reading Scale Alpha 2, adapted to Individual Testing," *Teachers College Record*, 18: 253–260, May, 1917.

McCall, W. A. — "A Uniform Method of Scale Construction," *Teachers College Record*, 22: 31–51, Jan., 1921.

Monroe, W. S. — *A Critical Study of Certain Silent Reading Tests*, University of Illinois Bulletin no. 8, vol. 19, no. 22, University of Illinois, Urbana; 52 pp., 1922.

Pressey, L. W., and Pressey, S. C. — "A Critical Study of the Concept of Silent Reading," *Journal of Educational Psychology*, 12: 25–32, Jan., 1921.

Ruch, G. M., and Current, W. F. — "Further Studies on the Reliability of Reading Tests," *Journal of Educational Psychology*, 17: 476–481, Oct., 1926.

Smith, W. A. — *The Reading Process*, ch. 11. The Macmillan Company, New York; 267 pp., 1922.

STONE, C. R. — *Silent and Oral Reading*, ch. 9. Houghton Mifflin Company, Boston; 306 pp., 1922.

— "Recent Developments in Silent Reading Tests," *Journal of Educational Research*, 6 : 102–115, Sept., 1922.

SYMONDS, P. M. — "The Accuracy of Certain Standard Tests for School Classification," *Journal of Educational Research*, 9 : 315–329, April, 1924.

THORNDIKE, E. L. — "An Improved Scale for Measuring Ability in Reading," *Teachers College Record*, 16 : 31–53, 17 : 40–67, Nov., 1915, Jan., 1916.

— "The Measurement of Achievement in Reading: Word Knowledge," *Teachers College Record*, 17 : 430–454, Nov., 1916.

— "The Measurement of Ability in Reading," *Teachers College Record*, 15 : 4, Sept., 1914.

— *The Teacher's Word Book*, Bureau of Publications, Teachers College, Columbia University, New York; 134 pp., 1921.

— "Word Knowledge in the Elementary School," *Teachers College Record* 22 : 334–370, Sept., 1921.

THORNDIKE, E. L., and SYMONDS, P. M. — "Difficulty, Reliability, and Grade Achievements in a Test of English Vocabulary," *Teachers College Record*, 24 : 438–445, Nov., 1923.

UHL, W. L. — "The Use of the Results of Reading Tests as a Basis for Planning Remedial Work," *Elementary School Journal*, 17 : 266–275, Dec., 1916.

ZIRBES, L. — "Diagnostic Measurement as a Basis for Procedure," *Elementary School Journal*, 18 : 505–522, March, 1918.

(Tests in Language)

ASHBAUGH, E. J. — "The Measurement of Language: What is Measured and Its Significance," *Journal of Educational Research*, 4 : 32–39, June, 1921.

BRIGGS, T. H. — "An English Form Test," *Teachers College Record*, 22 : 1–11, Jan., 1921; also TEACHERS COLLEGE BULLETIN, FOURTEENTH SERIES, NO. 6, Dec., 1922.

— "English Composition Scales in Use," *Teachers College Record*, 23 : 423–452, Nov., 1922.

CHARTERS, W. W. — "Constructing a Language and Grammar Scale," *Journal of Educational Research*, 1 : 249–257, April, 1920.

CHOU, H. H. C. — *The Measurement of Composition Ability*. Printed privately (probably 1922).

Courtis, S. A. — *Measurements of Classroom Products, Gary Public School Survey*, ch. 6. General Education Board, New York; 532 pp., 1919.

Dolch, E. W. — "Measurement of High School English," *Journal of Educational Research*, 4: 279–286, Nov., 1921.

Gunther, C. — "My Experience with the Hillegas Scale," *English Journal*, 535–542, Nov., 1919.

Hudelson, E. — "English Composition, Its Aims, Methods, and Measurement," Twenty-Second Yearbook of the National Society for the Study of Education, Part I; Public School Publishing Company, Bloomington, Ill., 1923.

— "Some Achievements in the Establishment of a Standard for the Measurement of English Composition in the Bloomington, Indiana, Schools," *The English Journal*, 5 : 590–597, Nov., 1916.

Hudelson, E., and Others — "Sixteen Spelling Scales," *Teachers College Record*, 21: 337–391, Sept., 1920.

Johnson, F. W. — "The Hillegas-Thorndike Scale for Measuring the Quality in English Composition by Young People," *School Review*, 21: 39–49, Jan., 1913.

Kayfetz, I. — "A Critical Study of the Hillegas Composition Scale," *Pedagogical Seminary*, 21 : 559–577, Dec., 1915.

Lewis, E. E. — "Scales for Measuring Special Types of English Composition," reviewed by M. J. Van Wagenen, *Journal of Educational Research*, 4: 427, Dec., 1921.

Strayer, G. D. — *Some Problems in City School Administration*, ch. 9. World Book Company, Yonkers; 234 pp., 1916.

Thorndike, E. L. — "Notes on the Significance and Use of the Hillegas Scale for Measuring the Quality of English Composition," *English Journal*, 2 : 551–561, Nov., 1913.

Trabue, M. R. — "Supplementing the Hillegas Scale," *Teachers College Record*, 18: 51, Jan., 1917; also Teachers College Bulletin no. 4, Tenth Series, Oct. 26, 1918.

Van Wagenen, M. J. — "Minnesota English Composition Scales: Their Derivation and Validity," *Educational Administration and Supervision*, 7: 481–499, Dec., 1921.

Willing, M. H. — *Valid Diagnosis in High School Composition*, Teachers College Contribution to Education, no. 230, Bureau of Publications, Teachers College, Columbia University, New York; 64 pp., 1926.

Wilson, G. M. — "Language Error Tests," *Journal of Educational Psychology*, 13: 341–349, 430–437, Sept., Oct., 1922.

CHAPTER VI

STANDARDIZED TESTS FOR THE TEACHER OF MATHEMATICS

In this chapter no reference will be made to tests of arithmetic. Arithmetic is often taught in high school, especially in commercial courses, but readers interested in knowing of the available tests in arithmetic may refer to manuals on testing in the elementary field, where they will find them much more adequately discussed than would be possible in the limits of this volume. Tests of elementary algebra and plane geometry will be described.

Elementary Algebra

OBJECTIVES OF INSTRUCTION

The objectives which are representative of the most progressive thought in algebra teaching may be summarized as follows:

1. Ability to read, understand, frame, evaluate, solve, and transform formulæ.
2. Ability with equations.
3. Ability to solve problems by the use of algebra.
4. Ability to read and construct graphs.
5. Ability to use and understand negative numbers.
6. Skill in algebraic technique.
7. Comprehension of general relationships.
8. Less well-defined objectives coming under the headings *disciplinary* and *cultural.*

There will be differences of opinion about the order of importance and about the extent to which each objective should be

pursued. This point has caused many of our existing standardized tests to be criticized on the basis that they test matters which are not objectives according to the best thought of to-day.

THE HOTZ FIRST-YEAR ALGEBRA SCALES[1]

These tests were constructed by Henry G. Hotz. For several years they have been representative of careful work in scale making in a high school subject. There are five separate scales designed to measure five phases of the work of first-year algebra: (1) Addition and Subtraction, (2) Multiplication and Division, (3) Equation and Formula, (4) Graphs, (5) Problems. There are two series, *A* and *B*. Series *B* is a longer scale in each case and contains all the exercises tried out by Hotz. Series *A* is a shorter scale and includes exercises equally spaced in difficulty so that each item is as much more difficult as the one preceding as that one is more difficult than the one preceding it. Series *A* is the one generally used because of its economy of time, although Series *B* is useful in case a more thorough testing of achievement is desired. The exercises in each scale are arranged in order of difficulty. The first exercise is usually so easy that a majority of high school pupils can solve it early in the course, whereas the last exercise is so difficult that many fail in attempting it even after a year's study. For instance, in the Equation and Formula Scale item, No. 1 is the equation $2x = 4$; while item No. 25 is the equation $\sqrt{x^2 - 1} - x = -1$. The Hotz scales were standardized on the basis of extensive testing. Eighty-four high schools in 11 states coöperated in giving the tests and over 16,000 pupils were tested in the process of standardization. The time allowance for Series *A* is 20 minutes for each scale, except for the problem scale and graph scale, which are allowed 25 minutes. This is usually sufficient time for each pupil to do all he is able to do.

[1] Published by Bureau of Publications, Teachers College, Columbia University (1918).

Time is not an important factor in giving the tests. The score is the number of exercises correctly solved. No partial credits are given and the key should be followed exactly as found in the Teacher's *Manual.*

Upton[1] states as his opinion that the Equation and Formula Scale is the one to select if time is available for only a single test. This one scale seems to cover more of the objectives of present-day algebra instruction and more important objectives than the other scales. The Problem Scale is perhaps next in importance. The Addition and Subtraction Scale and the Multiplication and Division Scale test technique or skill with the fundamental processes. Upton[2] suggests the following schedule of rotation in giving these scales:

1. At the end of three months	Addition and Subtraction Scale Equation and Formula Scale
2. At the end of six months	Multiplication and Division Scale Problem Scale
3. At the end of nine months	Equation and Formula Scale (repeated) Graph Scale

There are no duplicate forms at present, which is one disadvantage in using the Hotz Scales. We know that repetition of a test results in a practice effect or increase of score, owing to familiarity with the test itself, so that one has to be cautious in interpreting scores on a repetition of the test.

Norms are available at the end of three, six, and nine months of instruction, and are included in the *Teacher's Manual for First-Year Algebra Scales* by H. G. Hotz. Because comparisons are

[1] Upton, C. B. — "Standardized Tests in Mathematics for Secondary Schools."
[2] *Op. cit.*, p. 349.

possible at these times it is probably best to give the tests at the end of three, six, and nine months of the school year.

TABLE 31

HOTZ ALGEBRA SCALES: NORMS

SERIES *A*

Test	3 Months	6 Months	9 Months
Addition and Subtraction	5.0	6.8	7.9
Multiplication and Division	5.3	6.3	7.9
Equation and Formula	4.9	7.1	7.8
Problem	4.3	4.9	5.6
Graph	2.8 (4½ mos.)		5.6

SERIES *B*

Test	3 Months	6 Months	9 Months
Addition and Subtraction	9.7	12.9	14.4
Multiplication and Division	9.6	14.0	16.3
Equation and Formula	7.8	14.3	16.0
Problem	5.4	6.5	7.5
Graph	3.7 (4½ mos.)		7.2

The tests are easy to give and the directions for giving are simple.

Scales which increase in difficulty by approximately equal steps furnish a most reliable objective means for determining the actual achievement of a student or a class of students. It is well to keep in mind that a low average class score is not always, nor even quite generally, due to poor instruction. Any one of several causes may be operating to keep a class score down. It is the duty of the teacher, however, to study these causes and to learn which ones are affecting the efficiency of the instruction, in order that proper remedial measures may be applied. This is possible with a much

greater degree of certainty when the teacher knows the actual standard of achievement a class has attained. Such knowledge furnishes the teacher with a fact basis upon which to proceed and a motive with which to operate.

The extent of progress made by a class can be quite scientifically measured by submitting the same scale to a class at intervals of about three months. Teachers should be cautioned very specifically not to do any drill work upon the exercises or problems appearing in the scales. The most desirable method of measuring progress would be to have another parallel series of scales similar and equal in difficulty to Series *A*. It is hoped that such a series will soon be constructed. For diagnostic purposes the scales of Series *B* have been found to be more serviceable.

Finally, it must be stated emphatically that these are primarily *power* tests and with the time limits as now fixed, they are speed tests to a limited extent only, *i.e.*, the speed element is secondary. If a pure speed test is desired, the tests devised by Rugg and Clark could be used to supplement these tests. The latter would be particularly useful in determining whether a class has had sufficient drill upon the fundamentals.[1]

Of the Hotz Scales Thorndike[2] has some significant things to say which, by the way, have a much wider application.

The graduation (of these scales) is empirical, difficulty being measured by the percentage of a group of pupils who fail task by task. This difficulty is thus a function of the nature of the training had by the pupils of certain communities at a certain date (1917, 1918), as well as of the intrinsic essential difficulty of the task. The order of difficulty may be, and probably will be, greatly altered by such differences in the teaching of algebra as each new decade shows. For example, at the time and in the places where Hotz standardized these tests No. 17 below was a much harder task than No. 14;

14. $3m + 7n = 34$
    $7m + 8n = 46$
17. In the formula $RM = EI$, find the value of M

[1] HOTZ, H. G. — *First-Year Algebra Scales.*
[2] THORNDIKE, E. L. — *Psychology of Algebra*, pp. 176–179, 1923.

but it would almost certainly be the easier of the two for a group who had been taught the content of Nunn's *Exercises*, Book I, by the methods which Nunn advocates.

Any empirical gradation as to difficulty must, of course, be thus a function of the nature of the teaching in vogue; and be itself subject to change. The graded or "ladder" form of test has, none the less, great merits. First, it helps us to state how hard a task a class or an individual can perform with substantial mastery. Suppose, for example, that in the test for addition and subtraction the per cents correct for the class were:

Task	Per Cent
1	95
3	90
5	100
9	95
10	95
12	100
13	95
15	80
18	75
22	65
23	45
24	20

It becomes at once clear that this class has a substantial mastery of addition and subtraction through difficulty No. 13. There is a great difference between knowing $\frac{n}{2}$ things perfectly and half knowing n things. The graded or "ladder" test informs us concerning how many things a class knows and which they are.

Second, it measures with fairness and economy of time, over a wide range of ability. Individual differences in the capacities of pupils to learn algebra are such that if the pupils spend approximately the same time, say 400 hours, in study and class-work, some should have acquired far more ability than others. If ten tasks of about equal difficulty are set as the means of measuring them, many of them should do all ten perfectly and many of them should fail on all ten. The ladder test avoids "such undistributed perfects" and "undistributed zeros."

Third, it helps us to distinguish lack of ability from lack of care. Suppose, for example, that John and William have taken the Hotz test for addition

and subtraction, extended to include five tasks at each level of difficulty, with the following results:

Level of Difficulty	Number Right John	William
1	5	5
3	3	5
5	4	5
9	4	5
10	5	5
12	4	5
13	5	0
15	0	0
18	0	0
22	0	0
23	0	0
24	0	0
Total Right	30	30

John is obviously much more careless than William, who is indeed an Admirable Crichton in this subject. Carelessness is the failing in tasks easier than those one succeeds with and in tasks which one usually succeeds with. Both features are important. Graded tests specially reveal the former as the Rugg-Clark tests do the latter.

. Some features of the Hotz tests may be criticized on the ground that the situations involved are unreal and unlikely ever to be met with by a sane person in a real world. For example:

$$\frac{3-2x}{(x-1)^3}+\frac{x+1}{(x-1)^2}-\frac{1}{x-1}=$$

$$\frac{p^2+4p-45}{p^2+2p+4}\cdot\frac{p^3-8}{p^2-81}\cdot\frac{1}{3pr-15pr}=$$

$$\sqrt{x^2-1}-x=-1$$

"A gold watch is worth ten times as much as a silver watch, and both together are worth $132. How much is each worth?

"The total number of circus tickets sold was 836. The number of tickets sold to adults was 136 less than twice the number of children's tickets. How many were sold of each?

"The area of a square is equal to that of a rectangle. The base of the rectangle is 12 feet longer and its altitude 4 feet shorter than the side of the square. Find the dimensions of both figures."

THE RUGG-CLARK TESTS IN FIRST-YEAR ALGEBRA

These well-constructed tests differ markedly in principle from the Hotz tests just described. The Hotz tests are *power* tests affording information as to how difficult an exercise a pupil can do — the Rugg-Clark tests are *speed* tests affording information as to how much skill a pupil has acquired on tests of a given level of difficulty. The authors are Harold O. Rugg and John R. Clark. The work of standardization of these tests was done in Illinois in and about the vicinity of Chicago.[1] They come in two booklets, the first containing tests 1 to 9, and the second tests 10 to 14. The different tests are for the operations shown in Table 32 on page 121. These operations are representative of those most frequently taught in first-year algebra as found from a study of representative textbooks in use at the time the authors made their study. As textbooks change and new emphases occur in the algebra course some of the above tests naturally will have less value. Two other tests, No. 15, graphs, and No. 16, quadratic equations, were originally included, but have been discarded.

There are also two preliminary tests, *A* and *B*, which were intended to be given on the day preceding the test proper in order to familiarize the pupil with the procedure of testing under time conditions, the routine of passing papers, the starting and stopping of the test, and the taking of directions orally, as no directions are printed on the test booklets. Test *A* is a test in removing parentheses, and *B* a test in simple equations. The Rugg-Clark tests have been constructed on a cyclical basis. For example, in test five, numbers 1, 6, 11, 16, 21, etc., resemble each other in form, and differ only in the particular digits used. Likewise example No. 2 resembles No. 7 in form. This cyclical arrangement insures that the test as a whole remain of constant difficulty.

[1] Published by the University of Chicago Book Store, 5802 Ellis Avenue, Chicago, Illinois (1917).

The cyclical arrangement gives the tests high diagnostic power. Errors are easily located and recurring errors may be distinguished from accidental errors. A recurring error is one which will occur every time in the corresponding place in a new cycle. An acci-

TABLE 32

RUGG-CLARK ALGEBRA TESTS: TESTS, TIME ALLOWANCE, AND ILLUSTRATIVE EXAMPLES

No. of Test	Subject	Time	Illustrative Example To Be Written on Blackboard: Example	Answer
			Book I	
1	Collecting terms	4′	$7x + 5y - 4x - 3y$	$3x + 2y$
2	Substitution	4′	If $x = 2$ and $y = 3$, what does $2x^2 + 3xy = ?$	$8 + 18 = 26$
3	Subtraction	4′	From $2a + 3b - 4$ take $5a - 2b + 1$	$-3a + 5b - 5$
4	Simple equations	4′	$2x = 12 + 4x$	$x = -6$
5	Parentheses	2′	$3(4x - 7)$	$12x - 21$
6	Special products	3′	$(3a - 4)^2$	$9a^2 - 24a + 16$
7	Exponents	2′	$a^2 \cdot a^3$	a^5
8	Factoring	4′	$x^2 - 5x + 6$	$(x - 2)(x - 3)$
9	Clearing fractions	5′	$\frac{x+3}{4} - \frac{x-2}{5} = 0$	$5x + 15 - 4x + 8 = 0$
			Book II	
10	Fractional equations	9′	$\frac{2x+3}{5} - \frac{x-2}{3} = 1$	$x = -4$
11	Practical formulæ	5′	If $a = \frac{b}{c}$, what does $c = ?$	$c = \frac{b}{a}$
12	Quadratic equations	7′	$x^2 - 9x + 20 = 0$	$x = 4$ or 5
13	Simultaneous equations	12′	$2x + 3y = 5$ $4x - 3y = 1$	$x = 1$ $y = 1$
14	Radicals	3′	$\sqrt{a^2b^3}$	$ab\sqrt{b}$
			Supplementary Tests	
15	Graphs	12′		
16	Quadratic equations (irrational roots)	12′	$2x^2 + 5x - 1 = 0$	$x = \frac{-5 \pm \sqrt{33}}{4}$

dental error is one which is a solitary lapse and which does not repeat its appearance in the corresponding place in the previous or succeeding cycles.

In giving the test an illustrative example is to be written on the blackboard with its correct answer. The two booklets have been so arranged that each may be conveniently given during a class period.

Both the number attempted and the number right are used in scoring, and norms are given for both as follows. We are not told at what time of the year these norms apply:

TABLE 33

RUGG-CLARK ALGEBRA TESTS: NORMS

NO. OF PROBLEMS ATTEMPTED

	Test Number										
	2	5	6	7	8	10	11	12	13	14	15
Score of most efficient school	4.2	14.5	6.3	13.4	4.6	3.0	4.2	3.5	1.1	7.3	1.4
9 most efficient schools	3.5	11.6	5.8	12.5	4.1	1.5	3.2	2.9	1.0	5.8	2.5
Ave. of 27 schools	3.0	10.4	4.9	11.2	3.6	1.1	2.7	2.4	0.8	4.9	3.1
Score of poorest school	2.2	6.5	2.8	7.9	2.4	0.5	1.5	1.4	0.5	2.9	

NO. OF PROBLEMS RIGHT

	Test Number										
	2	5	6	7	8	10	11	12	13	14	15
Score of most efficient school	3.2	13.2	5.4	11.8	4.1	1.4	2.8	2.8	0.9	5.3	2.0
9 most efficient schools	2.7	11.0	4.8	10.7	3.5	0.8	1.6	2.0	0.7	3.9	4.2
Ave. of 27 schools	2.2	9.7	3.8	9.2	2.9	0.5	1.1	1.4	0.6	2.8	7.9
Score of poorest school	1.2	6.1	2.2	4.3	1.7	0.2	0.1	0.5	0.3	0.1	

These two kinds of scores need not be combined. The number of problems attempted is probably the better index of speed. The ratio of number of problems right to number of problems attempted expressed as a percentage is a measure of accuracy. Accuracy is more important than speed in algebra. One hundred per cent accuracy is the standard and the teacher ought to strive for 100% accuracy and to accept little less than that. However, speed is in certain ways a sign of mastery, so that speed should not be entirely sacrificed to accuracy. On the other hand, the teacher should guard against having pupils try many more examples than they can get right. The results of tests show that this practice is only too common. It is evidence that bad habits of work certainly kill off any of the valuable disciplinary value of accuracy supposed to accrue from a study of algebra. I wish to emphasize the importance of accurate timing in the giving of the Rugg-Clark tests. The validity of the results depends largely on the precision with which the work is started and stopped.

The Rugg-Clark tests do not come in duplicate forms, but such forms could easily be constructed.

Concerning the Rugg-Clark tests Thorndike[1] says:

> Another general principle is suggested by the directions, and by the titles which serve as directions. The Rugg-Clark tests measure a compound made up (*a*) of certain abilities to operate, and (*b*) of certain trade secrets of algebra. Such are in Test 1, the understanding of "Collecting terms"; in Test 4, knowledge that "Answer to $2x + 3 = 11$" means "Find what x equals"; knowledge that "·" means multiply; in Test 11 knowledge that "Solve for h" means "Find what h equals"; in Test 13 knowledge that when *Simultaneous Equations* is at the top of the page, the "answer" to $\begin{matrix} 2x + 3y = 10 \\ 3x - 2y = 1 \end{matrix}$ is the value of x and the value of y which will make both equations true statements.
>
> In general it seems desirable to keep measures of operating ability as free

[1] *Op. cit.*, pp. 172–173.

as possible from measures of knowledge of these trade secrets. We should all agree, for example, that Test 11, if put as:

$P = ahw$. Change the subject of the formula to h.

$C = \frac{E}{R}$. Change the subject of the formula to R, etc., would baffle many pupils who nevertheless had mastery of the operating. Yet "Solve for" is as truly a trade secret as "Change the subject of the formula to," though it is much more widely known. In practice for mastery, of course, the claims of brevity justify terms like "simplify" or "solve," and the assumption that the pupils will know what to do with, say, simultaneous equations. In tests of mastery, however, where brevity is far less important than clearness, and where pupils are to be compared who may have been taught different customs as to terms and procedures, it seems better, as said above, to test the power of operating separately from the power of understanding what operation is desired.

It will be understood that the Hotz tests and the Rugg-Clark tests are by no means competitors. They do not have the same function or purpose. The Hotz tests give measures of achievement or of progress. The Rugg-Clark tests, on the other hand, only indirectly measure progress. They are eminently diagnostic tests. They inform one as to the skill reached in various levels or processes in algebra and enable one to detect the weak spots in instruction or in a particular pupil's performance. The Rugg-Clark tests are more useful in connection with everyday instruction. The Hotz tests are more useful in measuring achievement or progress at the end of a term.

Mathematicians criticise these standardized tests in algebra very severely. David Eugene Smith[1] has objected to them on the grounds that

"the material required for testing the mechanical processes is often such as should play only a minor rôle, if any, in the education of the average citizen, and that the tests represent generally a dead level of dull grind, offering to the teacher only this ideal of an algebra course. He may escape from the curriculum making his own course; he may and should select

[1] SMITH, D. E. — "On Improving Algebra Tests," *Teachers College Record*, 24: 87–94, March, 1923.

from the textbook that which he needs for carrying out this plan; but he cannot escape from teaching those things that are required by outside examinations whether they be set by boards of regents, by colleges, or by educational testers."

Until mathematicians are agreed as to what should be the content of the elementary algebra course, tests will always be subject to similar criticism. The mathematics teacher resents being tested on the product of day-by-day teaching. She believes that the worthwhile outcomes are something more indefinite and intangible than the habits and skills acquired in actual teaching. But until the air clears and these indefinite objectives are given more solid substance in the way of definite psychological outcomes, we can do no better than test the apparent things that are taught according to a consensus of present practice. That is what the two tests described above do, more or less successfully.

Geometry

OBJECTIVES OF INSTRUCTION

If there is some confusion as to the objectives of algebra teaching, there is chaos concerning geometry. In general, geometry in school may be divided into two kinds — intuitive and demonstrative. Intuitive geometry comprises:

1. Ability to measure length and angles.
2. Knowledge of the methods of computing perimeters, areas, and volumes.
3. Skill in numerical computation involving perimeters, areas, and volumes.
4. Drawing to scale and the use of squared paper.
5. Geometric constructions.
6. Familiarity with simple geometric forms and their names.
7. And perhaps other abilities.

So far as the writer knows, there are tests for none of these objectives, although it is quite feasible to construct them. Before they can be constructed, however, one must know the most important terminology, forms, constructions, and formulæ. This, as a field of research, has not been opened up.

Demonstrative geometry has occupied a more recognized position in the high school curriculum and hence we have a few standardized tests, but this study has presented unusual difficulty in the way of test construction. Most standardized tests measure knowledge or information or a particular skill. But geometry is a reasoning process and the testing attack cannot be so straightforward. The attack must rest upon an analysis of the processes involved in the solution of a geometry original, a geometry construction, or a numerical problem in geometry. For this reason the tests that we now have in geometry have peculiar significance, because they show the way toward the testing of features of the reasoning process in any field.

MINNICK GEOMETRY TESTS

I cannot pass without mentioning the pioneer work of J. H. Minnick in this field. Minnick was not the first to construct geometry tests, but his were the first to be widely used. The tests, five in number, were published in 1918 and are described in his thesis, *An Investigation of Certain Abilities Fundamental to the Study of Geometry.*

The four fundamental abilities with which his study is concerned are:

1. The ability to draw a figure for a theorem.
2. The ability to state the hypothesis and conclusion accurately in terms of the figure.
3. The ability to recall additional known facts concerning the figure.

4. The ability to select from all the available facts those necessary for a proof and to arrange them so as to arrive at the desired conclusion.

Corresponding to these four abilities are four tests. Test *A* consists of a series of geometrical exercises for which the pupil is to draw the figure. In test *B* the theorem is stated and the figure is drawn; the pupil is to state the hypothesis and conclusion. In test *C* the pupil is asked to state as many facts as possible about a given figure. Test *D* consists of a series of exercises in which the figures are drawn, the hypothesis stated, and lists of known facts given; the pupil is asked to prove the exercises. The time limit for each is 30 minutes.

The limitation of these tests lies in the subjectivity of the scoring. This drawback greatly weakens their usefulness. As a matter of fact, Minnick has devised a system of scoring which is intended to facilitate diagnosis, although it greatly complicates the scoring. The tests could have been made more reliable by using the multiple choice form and extending the number of questions.

SCHORLING-SANFORD ACHIEVEMENT TEST IN PLANE GEOMETRY

The best standardized geometry test[1] is the "Schorling-Sanford Achievement Test in Plane Geometry." The original of this test was constructed by Raleigh Schorling, in 1921. It was then taken over by Miss Vera Sanford, who has put it through three revisions and constructed a second form of the last revision.

The test comes in five parts, as is shown in the following table. Each part is preceded by a page giving examples and in some cases trial exercises.

[1] Published by Bureau of Publications, Teachers College, Columbia University

TABLE 34

SCHORLING-SANFORD GEOMETRY TEST: TESTS, POSSIBLE SCORE, AND TIME ALLOWANCE

TEST	TOTAL POSSIBLE SCORE	TIME MINUTES
1 Completion Sentences . .	12	8
2 Conclusions from given data	12	12
3 Judging conclusions . . .	12	10
4 Analyzing constructions . .	12	10
5 Computation	12	12
	60	52

Norms are given for the end of the year as follows:

TABLE 35

SCHORLING-SANFORD GEOMETRY TEST: NORMS (JUNE)

	FORM A	FORM B
75 percentile	37.0	36.8
Median	30.6	31.2
25 percentile	25.0	27.5

REFERENCES

(Objectives in Mathematics)

NATIONAL COMMITTEE ON MATHEMATICAL REQUIREMENTS. — *The Reorganization of Mathematics in Secondary Education*, chs. 1–4, 1923.

RUGG, H. O., and CLARK, J. R. — *Scientific Method in the Reconstruction of Ninth Grade Mathematics*, SUPPLEMENTARY EDUCATIONAL MONOGRAPHS, VOL. 2, NO. 1, 1918.

THORNDIKE, E. L. — *The Psychology of Algebra. Ref.* Ch. 3.

(Tests)

CAWL, F. R. — "Practical Uses of an Algebra Standard Scale," *School and Society*, 10 : 88–90, July, 1919.

DOUGLASS, H. R. — *The Derivation and Standardization of a Series of Diagnostic Tests in the Fundamentals of First-Year Algebra*, UNIVERSITY OF OREGON PUBLICATIONS, VOL. 1, NO. 8, University of Oregon, April, 1921.

— *The Douglass Standard Diagnostic Tests for Measuring Achievement in First-Year Algebra — Revisions and Extensions*, UNIVERSITY OF OREGON PUBLICATIONS, VOL. 2, NO. 5. University of Oregon, 1924; 37 pp.

— "A Series of Standardized Diagnostic Tests in the Fundamentals of Elementary Algebra," *Journal of Educational Research*, 3: 396–403, Dec., 1921.

HARRIS, E., and BREED, F. S. — "Comparative Validity of the Hotz Scales and the Rugg-Clark Tests in Algebra," *Journal of Educational Research*, 6: 393–411, Dec., 1922.

HARRIS, E. — *A Study of the Hotz First-Year Algebra Scales and the Rugg-Clark Standard Algebra Tests.* University of Chicago Press, Chicago, 1919.

HOTZ, H. G. — *First-Year Algebra Scales*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 90; Bureau of Publications, Teachers College, Columbia University, New York, 1918.

— *Manual of Directions for First-Year Algebra Scales*, Bureau of Publications, Teachers College, Columbia University, New York, 1922.

MINNICK, J. H. — *An Investigation of Certain Abilities Fundamental to the Study of Geometry.* University of Pennsylvania, Philadelphia, 1918.

— "A Scale for Measuring Pupil's Ability to Demonstrate Geometrical Theorems," *School Review*, 27 : 101–109, Feb., 1919.

RUGG, H. O. — "Experimental Determination of Standards in First-Year Algebra," *School Review*, 24 : 37–66, Jan., 1916.

RUGG, H. O., and CLARK, J. R. — "The Improvement of Ability in the Use of the Formal Operations of Algebra by Means of French Practice Exercises," *School Review*, 25 : 546–554, Oct., 1917.

— "Standardized Tests and the Improvement of Teaching in First-Year Algebra," *School Review*, 25 : 113–132, Part I, Feb., 1917; 25 : 196–213, Part II, March, 1917.

SANFORD, V. — "A New Type Final Geometry Examination," *The Mathematics Teacher*, 18 : 22–36, Jan., 1925.

SMITH, D. E. — "On Improving Algebra Tests," *Teachers College Record*, 24 : 87–94, March, 1923.

UPTON, C. B. — "Standardized Tests in Mathematics for Secondary Schools," ch. 13 in *The Reorganization of Mathematics in Secondary Education*, 1923.

CHAPTER VII

STANDARDIZED TESTS FOR THE TEACHER OF SCIENCE

Science in high school suffers from a malady that is common to almost all the high school subjects. Science teachers are striving for far-away, long-range objectives which can only be reached by means of immediate short range objectives. Such objectives as abiding interests, appreciation, scientific attitude, or health are stated as the aims of science courses, whereas the day-to-day work comprises the learning of specific information, solving specific problems, performing specific experiments, and the like. These latter functions yield easily to measurement — the less tangible functions, such as interests and attitudes, are stubborn problems for the educational scientist.

GENERAL SCIENCE

OBJECTIVES OF INSTRUCTION

The objectives of general science have had many statements, but the following by Curtis[1] perhaps satisfy by being succinct and definite.

1. To give extensive and serviceable knowledge of scientific facts and principles.
2. To broaden, quicken, create, and satisfy in the individual scientific interests.
3. To train for the worthy use of leisure.
4. To develop scientific attitudes.

[1] CURTIS, F. D. — *Some Values Derived from Extensive Reading of General Science*, TEACHERS COLLEGE CONTRIBUTION TO EDUCATION, NO. 163, P. 1, 1924.

Naturally the first of these — acquisition of knowledge — is the most definite and easiest to test. The development of scientific attitudes is also capable of being tested. The development of interests cannot be directly tested, but may be estimated by means of appropriate questionnaires. Whether one is trained for effective use of leisure time depends largely on the other objectives.

RUCH-POPENOE GENERAL SCIENCE TEST

A satisfactory test for measuring the information side of general science is the "Ruch-Popenoe General Science Test."[1] It is a revision of an earlier test, known as the "Range of Information Test in General Science," described by Ruch[2] first in 1919. This was the beginning of Ruch's fruitful work in test construction. The work was carried from Oregon to California (Stanford), where Ruch and Popenoe reworked the test into its present shape. It is designed for use in the 8th and 9th grades — the grades where general science is usually taught. The test comes in two forms, *A* and *B*, identical in construction and equated for difficulty.

The test is divided into two parts.

Part I is composed of fifty items of general information concerning familiar elementary scientific facts, principles, concepts, terms, definitions, and applications. These fifty items sample a wide range of relatively simple knowledge in the fields of physics, chemistry, astronomy, agriculture, botany, zoölogy, and physiology.

Part II measures the ability of the pupil to identify apparatus, organisms, structures, and principles, and to apply principles of science to the solution of simple problems. For this purpose twenty diagrams and drawings have been prepared, together with completion exercises based upon these illustrations.[3]

[1] The test was published in 1923 by the World Book Company.

[2] Ruch, G. M. — "A Range of Information Test in General Science," *General Science Quarterly*, 4 : 257–262, Nov., 1919; 5 : 15–19, Nov., 1920.

[3] From Ruch-Popenoe General Science Test: Manual of Directions. Copyright 1923 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

The subject matter of the test has been analyzed as follows:

1. Biological sciences (botany, physiology, zoölogy, etc.) 30%
2. Chemistry 12%
3. Physics and mechanical applications 38%
4. Earth sciences (agriculture, astronomy, physiography, etc.) . . 20%

The manual of directions contains explicit directions for giving the test. The score equals the sum of the correct items in Part I plus one half the correct items in Part II. A key is provided.

TABLE 36

RUCH-POPENOE GENERAL SCIENCE TEST: TESTS, TIME ALLOWANCE, AND POSSIBLE SCORE

	No. Items	Time in Minutes	Maximum Score
Part I .	50	15	50
Part II .	80	25	40
		40	90

Norms based on the testing of 363 pupils are given in the manual for the end of the year as follows:

TABLE 37

RUCH-POPENOE GENERAL SCIENCE TEST: NORMS (JUNE)

10% of the pupils reach or exceed	53.4
25% of the pupils reach or exceed	44.6
50% of the pupils reach or exceed	34.7
75% of the pupils reach or exceed	24.9
90% of the pupils reach or exceed	18.8

A teacher might well criticize the validity of the Ruch-Popenoe test for use in his own class. This test naturally does not fit any

one text or any one course of study. There are questions in the test for many topics which a teacher will not have taken up in class. It is an open question as to whether this feature hurts the value of the test. Probably it lowers the reliability slightly. Perhaps one should consider the test as made up of only those items which cover topics taken up specifically in any one class, and consider the rest of the questions as dead wood, whose presence in no way affects the results of the test.

DVORAK GENERAL SCIENCE SCALES

Only brief mention can be made of the "Dvorak General Science Scales," which are well-constructed tests.[1] One form of the Dvorak Scale, *R–1*, may be used as a prognosis test before the pupil has taken the general science course, — two other forms, *S–2* and *T–2*, are to measure achievement in the general science course. These tests are scaled and standardized on the system devised by Van Wagenen. Curtis,[1] in his experiments on the value of extensive reading in general science, preferred the Dvorak tests because of "their careful standardization, ease of administering and scoring, and all-round excellence."

BIOLOGY

OBJECTIVES OF INSTRUCTION

The course in biology in high school has the following objectives:

1. A first-hand acquaintance with biological facts, phenomena, and principles.
2. Knowledge of the principles of hygiene and sanitation.
3. Ability to solve problems using biological principles.
4. An appreciation of the importance and value of knowledge of biology in contemporary life.

[1] *Op. cit.*

RUCH-COSSMAN BIOLOGY TEST

The "Ruch-Cossman Biology Test"[1] is the most satisfactory standardized test of biology that we have to day. It measures outcomes of the biology course that come under the first of the objectives mentioned above, and, to a limited degree, the third objective.

The authors are G. M. Ruch and Leo H. Cossman. This test was started in 1918, although the tests were only published in 1924. They are provided in two duplicate forms and are useful in the course in biology. The examination consists of five separate tests as follows:

TABLE 38

RUCH-COSSMAN BIOLOGY TEST: TESTS, NUMBER OF ITEMS, AND TIME ALLOWANCE

TESTS	NO. ITEMS	TIME IN MINUTES
1. General biological information .	40	12
2. "Best answer" statements . . .	18	8
3. Identification of structures from drawings	15	5
4. Laws of Mendelian inheritance . .	4	3
5. Completion exercises	35	10
Total	112	38

One of the most careful pieces of work in selecting valid items for a test was done by Cossman in the construction of this test. The scoring is simple. The maximum total score is 116 points, the items on test 4 each counting two.

The norms given in the manual for the end of the second semester are:

[1] Published by the World Book Company.

TABLE 39

RUCH-COSSMAN BIOLOGY TEST: TENTATIVE PERCENTILE NORMS

10% of the students reach or exceed . . .	56
20% of the students reach or exceed	50
25% of the students reach or exceed	47 (upper quartile)
30% of the students reach or exceed	45
40% of the students reach or exceed	41
50% of the students reach or exceed	37 (median)
60% of the students reach or exceed	33
70% of the students reach or exceed	31
75% of the students reach or exceed	29 (lower quartile)
80% of the students reach or exceed	27
90% of the students reach or exceed	22

PHYSICS AND CHEMISTRY

OBJECTIVES OF INSTRUCTION IN PHYSICS

The objectives of instruction in physics are similar to those of the other sciences, and may be summarized briefly as follows:

1. Knowledge of facts, phenomena, and principles of physics.
2. Ability to solve problems involving laws and principles of physics.
3. Skill in the technique of the physics laboratory.
   (*a*) Accurate observation
   (*b*) Methodical attack on a problem
   (*c*) Habits of order and neatness with materials
   (*d*) Technique of manipulation of apparatus
   (*e*) Ability to measure
   (*f*) Ability to calculate
   (*g*) Ability to arrange and tabulate data

IOWA PHYSICS TESTS

At the present time the most carefully constructed physics tests are those of Harold L. Camp. Camp originally tried out

435 questions in physics from which scales were constructed — that is, the difficulty value of each question was determined. From these scales tests were constructed known as the "Iowa Physics Tests."[1] The tests come in three separate sections — Mechanics, Heat, and Electricity and Magnetism. When the Public School Publishing Company took them over, they were expanded so as to include two forms, equated in difficulty on the basis of Camp's original scale values. Each separate test is designed to be given following the completion of the section in physics to which it is devoted.

Camp[2] states that the items of his tests measure either (1) knowledge of the fundamental principles or (2) ability to put the knowledge to use in solving problems one meets in ordinary life. As to the precise selection of material, Camp[3] says:

> Dr. Daniel Starch made a careful analysis of five textbooks (Black and Davis, Carhart and Chute, Hoadley, Mann and Twiss, and Millikan and Gale) most commonly used for high school physics, and found 102 principles common to all five of them. These results were used as a guide in formulating test exercises.

The exercises are all of the problem or single-answer type. Originally the material was "arranged into 47 preliminary tests of nine or ten exercises each. Correct solutions of these exercises were attempted by approximately 3,500 boys and girls in 129 of the best high schools in Iowa. Each exercise was attempted by an average of 99 different pupils (minimum 60) and each test was given from eight to ten weeks after the work in that particular phase of physics had been finished."[4]

One novel feature of this preliminary work was that the time

[1] Published by the Public School Publishing Company, Bloomington, Ill.; originally by the Extension Division of the University of Iowa in 1920.

[2] CAMP, H. L. — *Scales for Measuring Results of Physics Teaching*, IOWA STUDIES IN EDUCATION, VOL. 2, NO. 2, p. 11, 1921.

[3] *Op. cit.*, p. 12.

[4] CAMP. H. L. — *Op. cit.*, p. 12.

necessary to solve each problem was determined. From these data it was easy to put together a test of the right length to occupy a school period. The data concerning each test are:

TABLE 40

IOWA PHYSICS TESTS: TESTS, NUMBER OF ITEMS, AND TIME ALLOWANCE

TEST	NO. ITEMS	TIME
Mechanics . . .	12	45 minutes
Heat	11	45 minutes
Electricity and magnetism	11	40 minutes

One is instructed to score the test by obtaining the total of the scale values of the correctly answered items. This method of scoring is obsolete and is no longer employed. Many test makers have experimented with scoring by adding scale values and by simply adding correctly answered test items and have found the results about identical (the correlation is usually over + .95). The norms given below are in terms of these scale values.

TABLE 41

IOWA PHYSICS TEST: PERCENTILE NORMS

PER CENT	MECHANICS	HEAT	ELECTRICITY AND MAGNETISM
75 percentile . .	49.0	55.1	54.6
Median . . .	33.5	41.9	39.9
25 percentile . .	21.0	27.4	28.0

One may make several criticisms of this test. The form of the test could be improved. The information or definition type questions could be thrown into the multiple-response mold and

consequently made more objective. Another advantage that would accrue from this is that it would permit the inclusion of many more questions, thereby making the test more comprehensive. Problems are so time-consuming that they are "wasteful" in a test. That is, a test with only 11 or 12 items probably possesses low reliability. (Camp does not inform us as to the reliability.) One may analyze the problems into the constituent mental processes used in solving a physics problem and test these separately. For instance, item No. 7 in form 2 of the electricity and magnetism test is:

At 15 cents per kilowatt hour, what will it cost to run for 10 hours a 220-volt motor drawing a current of 25 amperes?

In doing this problem one has to

1. recall the formula or principle to be used;
2. substitute correctly in the formula, including the correct use of units;
3. perform the arithmetical operations correctly;
4. express the result in the correct units.

If 1, 2, and 4 are the steps in the process which specially concern the physics teacher, they should be tested separately. This analytic (and diagnostic) testing would make a physics test more reliable and also more comprehensive. One need not hesitate to mutilate or abridge a problem for the purposes of testing.

From Camp's own description of the determination of his scale, one wonders whether his scale values have been accurately determined. If some items were tried in the standardization by only sixty pupils and if these pupils happened to be in particularly good or poor groups, a not unwarranted assumption, the scale values in many cases might be far from the true values.

The method of scoring has already been criticized as awkward and cumbersome. A simple summation of correct items would be preferable, and norms should be provided in these units.

There are no specific tests of various outcomes of laboratory instruction.

OBJECTIVES OF INSTRUCTION IN CHEMISTRY

The outcomes of instruction in chemistry are of the same nature as for the other sciences and may be repeated as follows:

1. Knowledge of facts, phenomena, and principles of chemistry.
2. Ability to solve problems involving laws and principles of chemistry.
3. Skill in the technique of laboratory manipulation.
4. Less specific outcomes, such as attitudes, interest, habits of inquiry, accuracy, and certain appreciations of the importance and achievement of chemistry in contemporary life.

As in the case of the other sciences, the first two objectives, namely, the acquisition of facts and the ability to solve problems, are the abilities most readily tested.

POWERS GENERAL CHEMISTRY TEST

One of the best tests in chemistry is the "Powers General Chemistry Test,"[1] constructed by S. R. Powers. The experimental work for the test was begun in 1920, but in its present form it appeared in 1924. It comes in two equivalent forms, *A* and *B*, and is useful at any point in the course in chemistry.

The test consists of two parts. Part I is a test of the range of information about chemistry and includes 30 items constructed on the multiple-response form. Part II contains 37 items and tests ability to write formulas and equations, to give the chemical names of common substances, and to do simple calculations.

Powers[2] does not tell in his monograph that he selected his items on any basis other than difficulty. In all, he found the

[1] Published by the World Book Company.

[2] Powers, S. R. — *A Diagnostic Study of the Subject Matter of High School Chemistry.*

difficulty value of 350 items. His statement of his criteria of selection is:

There are at least three criteria of determining suitability of items. In the first place, a given item must test some ability (power) which it is desirable for all students to possess. . . . In the second place, the items must be of sufficient difficulty to make demands for ability. Obviously items which are answered correctly by all or nearly all students are not adequate for measuring achievement. Finally, the items must not be too difficult. The scales consist of items arranged in increasing order of difficulty. The first items are so easy that nearly all students will do them correctly. At the other extreme there are items so difficult that only a few of the superior students will be able to do them correctly. The increase in difficulty from very easy to very difficult items is by small and uniform steps.

No consideration of importance, frequency in school texts, or frequency in school examinations seems to have been employed. One must trust the wise selection of items to the wide experience and good judgment of Professor Powers.

The time allowance for either form is 35 minutes. The score is the sum of the correctly answered items. Powers originally used the scoring device perfected by Van Wagenen, but later abandoned this scoring method in favor of the simpler one of merely summing correct units, because the two methods correlated + .98.

Percentile norms are given in the following table:

TABLE 42

POWERS GENERAL CHEMISTRY TEST: PERCENTILE NORMS

A student making a score of 64 exceeds 98% of students
A student making a score of 55 exceeds 90% of students
A student making a score of 46 exceeds 75% of students
A student making a score of 36 exceeds 50% of students
A student making a score of 26 exceeds 25% of students
A student making a score of 17 exceeds 10% of students
A student making a score of 8 exceeds 2% of students

The test apparently is a good one. Objective data show that it agrees fairly well with teachers' marks in the course. The difficulty values have been very accurately determined in each instance on from 418 to 1209 cases. The order of difficulty remains quite constant from one year to the next year.

GERRY'S TEST OF HIGH SCHOOL CHEMISTRY

Another chemistry test having nearly equal merit with the Powers test is the "Gerry Test of High School Chemistry."[1] The basis for the selection of the original items was an analysis of the College Board questions from 1911 to 1920, checked by reference to four textbooks. Ninety-six questions were used for the experimental determination of difficulty, the multiple choice and completion forms being used. Seven hundred seventy pupils took the preliminary tests.

Gerry uses a novel and highly ingenious method of selecting items of apparently the same difficulty, that is, items answered correctly by the same proportion of pupils. His method was essentially to select that item which shows the highest percentage of pupils answering correctly all easier questions. For example, consider two items x and y answered correctly by 50% of a group of pupils. If 70% of pupils that answer x correctly answer all easier questions correctly, whereas 75% of pupils that answer y correctly answer all easier questions correctly, then y would be considered the better item. This procedure, although correct as far as it goes, shows a failure to grasp the whole problem and gives evidence of a lack of expert statistical advice in the construction of the test.[2]

[1] Published by the Harvard University Press, Cambridge, Mass.

[2] Difficulty of a question is not the prime consideration in determining the inclusion of an item in a test. The main criterion for including an item should be the correlation of the item with scores on the whole test; or better still, with an independent measure of the function being measured. This correlation may be computed by the bi-serial r formula. Kelley has pointed out that the problem is not

Gerry's test shows lower reliability than one would wish. This is in part due to the small number of items (25) finally included in each form. The test is intended to occupy the time of the ordinary high school period.

GLENN-OBURN PHYSICS AND GLENN-WELTON CHEMISTRY TESTS

One should not discuss the subject of standardized tests in science without mentioning the tests of Glenn and Oburn in physics and Glenn and Welton in chemistry. These are by far the most comprehensive tests to-day in any high school subject. In physics, for instance, the authors have constructed 26 objective tests covering all phases of the physics course, as may be seen by noting the list on page 143.

Four types of exercises have been used. Multiple choice (MC), one word (OW), true false (TF), and completion (C). Complete norms are provided for each test. The title of these tests, "New Types of High School Physics Tests for Investigational and Instructional Purposes," show what the tests are designed for. Glenn and his co-workers are using these tests in an investigation of the curriculum in physics and chemistry to determine exactly what are the outcomes of instruction. The tests are also useful for classroom testing. After covering a certain topic a teacher would spend 15–30 minutes in testing the class in that topic. The summary pages in each booklet provide excellent motivation. To get the greatest benefit from a test of this kind, remedial work should follow a test for those whose low scores show that they have failed to master a section. This remedial work would be impos-

quite as simple as this; that to make a complete study as to which items to include, one should determine all the intercorrelations between items and determine the partials. This is a rather heroic task to perform, but the work can be abbreviated by short cuts (see Kelley, *Statistical Methods*, Section 88). McCall and his students have done original work on this problem of criteria for the choice of individual items in a test. (See W. A. McCall and his students, *Teachers College Record*, 27: 394-415, Jan., 1926.)

TABLE 43

GLENN-OBURN PHYSICS TESTS: TESTS, TIME ALLOWANCE, AND NUMBER OF QUESTIONS

Test	Type	Time Minutes	Number of Questions
1. Measurement of length, mass, time, and density	MC	15	15
2. Liquid pressure: Pascal's law	OW	25	15
3. Air pressure; Pneumatic devices	OW	25	20
4. The molecular theory of matter	TF	20	40
5. Force and motion	C	30	25
6. Work, power, and energy	C	25	25
7. Mechanics of solids, liquids, and gases (Summary)	MC	30	25
8. Temperature measurements and expansion phenomena	OW	25	25
9. Transformations of work and heat energy	OW	25	20
10. The phenomena of melting and boiling	TF	25	50
11. Conduction, convection, and radiation phenomena	M	25	20
12. Heat (Summary)	C	25	25
13. Magnetic phenomena	MC	20	20
14. Static electrical phenomena	TF	25	50
15. Heating, chemical, and magnetic effects of the electric current	MC	20	20
16. Heating, chemical, and magnetic effects of the electric current	OW	20	25
17. The dynamo, motor, and transformer	C	25	25
18. Magnetism and electricity (Summary)	OW	20	20
19. The nature and transmission of sound	TF	20	25
20. The musical scale and musical instruments	C	25	20
21. Sound (Summary)	OW	20	15
22. The transmission of light	MC	20	15
23. Mirrors, lenses, and optical instruments	C	25	20
24. The phenomena of color	OW	25	25
25. Invisible radiations; wireless telegraphy and telephony	TF	20	40
26. Light (Summary)	C	20	30

sible, of course, under typical high school organization, but there is no reason why the organization should not in some way be altered to permit remedial work.

The list of tests in the chemistry series is as follows:

TABLE 44

GLENN-WELTON CHEMISTRY TESTS:
TESTS, TIME ALLOWANCE, AND NUMBER OF QUESTIONS

TEST	TYPE	TIME MINUTES	NUMBER QUESTIONS
1. Elements, Compounds, and Mixtures	TF	15	35
2. Oxygen, — Its Properties and Uses	OW	30	30
3. Hydrogen, — Its Properties and Uses	MC	30	20
4. The Metric System	C	30	35
5. Measurement of Gases	OW	30	20
6. Transformation of Matter and Energy	C	30	25
7. Water and Hydrogen Peroxide	C	30	25
8. Molecular and Atomic Weights: The Atomic Theory	TF	15	35
9. Formulas and Equations	C	40	25
10. The Chemical Symbols	OW	30	80
11. Carbon, Carbon Monoxide, and Carbon Dioxide	C	30	35
12. How Atoms Combine: Valence	C	30	45
13. Nitrogen and the Rare Gases in the Atmosphere	C	30	27
14. Chlorine and Hydrochloric Acid	OW	40	40
15. Sodium and Sodium Hydroxide	OW	30	45
16. Properties of Solutions: Electrolysis	MC	20	25
17. Compounds of Nitrogen	C	30	50
18. Types of Reactions	M	30	30
19. The Compounds of Sulfur	C	35	50
20. The Periodic Law	C	20	30
21. The Halogen Family of Elements	OW	40	50

TABLE 44 (*Continued*)

Test	Type	Time Minutes	Number Questions
22. Petroleum and the Hydrocarbons	OW	30	20
23. The Fuel Gases	C	25	30
24. Some Important Coal-tar Compounds	OW	30	30
25. Carbohydrates and Textile Products	OW	30	35
26. Alcohols, Organic Acids, and Foods	C	30	40
27. Phosphorus, Arsenic, Antimony, and Bismuth	C	30	45
28. Silicon and Boron	OW	30	30
29. Colloids	TF	20	40
30. The Sodium Family	OW	30	30
31. The Calcium Family	OW	30	35
32. Magnesium, Zinc, and Cadmium	OW	30	30
33. Aluminum and the Silicates . .	OW	30	30
34. Iron, Cobalt, and Nickel . . .	C	30	40
35. Copper, Mercury, Silver, Tin, and Lead	C	30	45
36. Manganese, Chromium, Gold, Platinum, Radium, and Tungsten	TF	20	50

All in all the sciences have among the best tests in the high school to-day. Perhaps this should be so. Perhaps the natural scientists, recognizing the importance of accurate measurements in their own field, accept more readily the use of more accurate measuring instruments in education. Also, the sciences, being bodies of organized knowledge, lend themselves more readily than some of the other subjects to testing. Whatever the reason, the sciences have led the way in the construction of standardized tests for the high school.

The "Michigan Botany Test"[1] by O. W. Laidlaw and Clifford Woody is a satisfactory test in botany.

[1] Published by the Public School Publishing Company.

REFERENCES

(Objectives)

CURTIS, F. D. — *Some Values Derived from Extensive Reading of General Science,* TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 163, Bureau of Publications, Teachers College, Columbia University, New York; 142 pp., 1924.

REORGANIZATION OF SCIENCE IN SECONDARY SCHOOLS — *A Report of the Commission on the Reorganization of Secondary Education,* appointed by the National Education Association; U. S. BUREAU OF EDUCATION, BULLETIN NO. 36, 1920.

(Tests)

CAMP, H. L. — "An Evaluation of Standard Tests and Suggested Uses in Improving Physics Teaching," *School Science and Mathematics,* 23 : 441–446, May, 1923.

— "Scales for Measuring Results of Physics Teaching," *Journal of Educational Research,* 5 : 400–405, May, 1922.

— "Scales for Measuring Results of Physics Teaching," *Iowa Studies in Education, Vol. 2, no. 2;* University of Iowa, Iowa City, 1921.

CURTIS, F. D. — See above under Objectives.

DVORAK, A. — *A Study of Subject Matter and Achievement in General Science;* monograph reprinted from *General Science Quarterly* 10: Nov., 1925, Jan., March, May, 1926, 100 pp.

GERRY, H. L. — "A Test of High School Chemistry," *Harvard Bulletin in Education, no. 9,* Harvard University Press, 1924.

POWERS, S. R. — "Achievement in High School Chemistry: An Examination of Subject Matter," *Teachers College Record,* 25 : 203–211, 1924; also *School Science and Mathematics,* 25 : 53–62, Jan., 1925.

— "The Achievement of High School and Freshman College Students in Chemistry," *School Science and Mathematics,* 21 : 366–377, April, 1921.

— "A Comparison of the Achievements of High School and University Students in Certain Tasks in Chemistry," *Journal of Educational Research,* 6 : 332–343, Nov., 1922.

— *A Diagnostic Study of the Subject Matter of High School Chemistry,* TEACHERS COLLEGE CONTRIBUTION TO EDUCATION, NO. 149, Bureau of Publications, Teachers College, Columbia University, New York; 84 pp., 1924.

POWERS, S. R. — "Tests of Achievement in Chemistry," *Journal of Chemical Education*, 1 : 139–144, 1924.

RUCH, G. M. — "A New Test in General Science," *General Science Quarterly*, 7 : 188–197, March, 1923.

— "A Range of Information Test in General Science," *General Science Quarterly*, 4 : 257–262, Nov., 1919; 5 : 15–19, Nov., 1920.

— "Tests and Measurements in High School Science," *School Science and Mathematics*, 23 : 885–891, Dec., 1923.

RUCH, G. M., and COSSMAN, L. H. — "Standardized Content in High School Biology," *Journal of Educational Psychology*, 15 : 285–296, May, 1924.

RUCH, G. M., and POPENOE, H. T. — "Measurement of Ability in General Science," *School Science and Mathematics*, 23 : 545–551, June, 1923.

(Scale scores *vs.* number items correct)

CHARTERS, W. W. — "Constructing a Language and Grammar Scale," *Journal of Educational Research*, 1 : 249–251, April, 1920.

DOUGLASS, H. R. and SPENCER, P. L. — "Is it Necessary to Weight Exercises in Standard Tests?" *Journal of Educational Psychology*, 14 : 109–112, Feb., 1923.

MONROE, W. S. — "The Description of the Performances of Pupils on Exercises of Varying Difficulty," *School and Society*, 15: 341–343, Mar. 15, 1922.

POWERS, S. R. — *A Diagnostic Study of the Subject Matter of High School Chemistry*, pp. 71–72. See ref. above.

CHAPTER VIII

STANDARDIZED TESTS FOR THE TEACHER OF LANGUAGE

LATIN

OBJECTIVES OF INSTRUCTION

The objectives of the study of Latin in high school have been set forth at great length in Part One of the *Classical Investigation*[1] and may be enumerated as follows:

1. Ability to read new Latin after the study of the language in school or college has ceased.
2. Increased ability to understand Latin words, phrases, abbreviations, and quotations occurring in English.
3. Increased ability to understand the exact meaning of English words derived directly or indirectly from Latin and increased accuracy in their use.
4. Increased ability to read English with correct understanding.
5. Increased ability to speak and write correct and effective English through training in adequate translation.
6. Increased ability to spell English words of Latin derivation.
7. Increased knowledge of the principles of English grammar and a consequently increased ability to speak and write grammatically correct English.
8. Increased ability to learn the technical and semi-technical terms of Latin origin employed in other school studies and in professions and vocations.

[1] American Classical League — *The Classical Investigation, Part I*, pp. 38-79.

9. Increased ability to learn other foreign languages.
10. The development of certain considerable habits and ideals which are subject to spread, such as habits of sustained attention, orderly procedure, overcoming obstacles, perseverance; ideals of achievement, accuracy, and thoroughness; and the cultivation of certain general attitudes, such as dissatisfaction with failure or with partial success.
11. Development of the habit of discovering identical elements in different situations and experiences and of making true generalizations.
12. The development of correct habits of reflective thinking applicable to the mastery of other subjects of study and to the solution of analogous problems in daily life.
13. Increased ability to make formal logical analyses.
14. Development of an historical perspective and of a general cultural background through an increased knowledge of facts relating to the life, history, institutions, mythology, and religion of the Romans; an increased appreciation of the influence of their civilization on the course of western civilization; and a broader understanding of social and political problems of to-day.
15. Increased ability to understand and appreciate references and allusions to the mythology, traditions, and history of the Greeks and Romans.
16. The development of right attitudes toward social situations.
17. A better acquaintance through the study of their writings with some of the chief personal characteristics of the authors read.
18. The development of an appreciation of the literary qualities of the Latin authors read and of a capacity for such appreciation in the literature of other languages.
19. A greater appreciation of the elements of literary technique employed in prose and verse.

20. Improvement in the literary quality of the pupil's written English.
21. An elementary knowledge of the simpler general principles of language structure.

For the Latin teacher to attempt to test all of these outcomes of instruction in the Latin class would be going rather far afield. Numbers 3, 4, 5, 6, and 7 could be tested by tests described in the chapter on English tests. The majority of the disciplinary and cultural objectives cannot be directly tested at the present time as there are no instruments available. The foregoing list of objectives illustrates well the dislike that secondary school people have of stating as objectives the results of day-to-day activity. They prefer to ignore the immediate and concentrate on the distant. So the immediate objectives, such as ability to translate, ability to give or place syntactical forms, etc., have been omitted from the list of objectives.

ULLMAN-KIRBY LATIN COMPREHENSION TEST

For testing the first objective in the list — ability to read Latin — there is an excellent test, the "Ullman-Kirby Latin Comprehension Test,"[1] constructed by B. L. Ullman and T. J. Kirby in 1922. This test was constructed for the American Classical League as an instrument to use in the *Classical Investigation.* It comes in two forms which are not strictly comparable. Data collected show that the means in the two forms for any semester are not more than one and one-half score units apart. Form I is slightly easier for second and eighth semester pupils than form II, but form II is slightly easier for fourth and sixth semester pupils than form I. The test is useful at any stage in the study of Latin. Pupils who have studied Latin only one year can make scores of from 5 to 13 points on the average.

[1] Published by the Extension Division, University of Iowa, Iowa City.

The test consists of ten paragraphs of Latin. There are two elementary paragraphs composed especially for the test, two selections from Cæsar, one from Cicero, one poetical selection, a third selection from Cæsar, a second from Cicero, a second poetical quotation, and a fourth selection from Cæsar. The authors state that the "stress on Cæsar was due not to the conviction that his works are necessarily the most important, but rather because elementary pupils can, under present conditions, read Cæsar more easily than other classical Latin." [1] On each of these paragraphs three or four questions are asked in English concerning material in the paragraph, and the pupil is expected to answer the questions in English. The following paragraph, used as a practice paragraph, illustrates the method.

Duōs fīliōs agricola habēbat, Mārcum et Sextum. Magnus erat Sextus et Mārcus erat parvus. Pater parvum fīlium magis amābat et eum ad urbem saepe mittēbat.

a. Which was more loved by his father, Sextus or Marcus?
b. Where did his father send Marcus?

Thirty minutes are allowed.

There are thirty-three questions on the test and the score is the number of correct answers. The test and scoring key are similar in form to the Thorndike-McCall reading tests and scoring keys. Correct answers or the essential portion of correct answers are printed in heavy block type in the key. Variant forms are placed in parentheses. Samples of wrong answers are also given as a guide to the scorer. The key does not include all possible correct answers and in many cases the scorer will have to use judgment. This means that the test should be scored by some one who is familiar with Latin.

The norms given on the class record sheet are:

[1] Ullman, B. L., and Kirby, T. J. — "A Latin Comprehension Test," *Journal of Educational Research,* 10: 308–317, Nov., 1924.

TABLE 45
ULLMAN-KIRBY LATIN COMPREHENSION TEST
SEMESTER NORMS

Semester	Form I			Form II		
	25 Percentile	Median	75 Percentile	25 Percentile	Median	75 Percentile
Second	8.1	10.5	13.0	5.5	8.0	10.9
Third	11.5	14.1	16.7	10.7	13.8	17.2
Fourth	14.7	17.5	20.6	13.9	17.5	19.9
Fifth	17.7	20.8	23.9	19.3	21.7	23.9
Sixth	18.7	22.1	25.1	19.3	22.2	25.0
Seventh	19.1	24.4	27.4	21.3	24.8	27.5
Eighth	21.6	25.3	28.4	21.4	25.1	27.9

The test is a very satisfactory one to give. It engages a class's attention and the pupils work well on it during the time allotted. There is always something on the test that every one can do. The norms show that there is a regular increase in the ability tested during the four years of high school. The test does not show as high reliability as might be desired. One cannot, however, question the validity of the test. It certainly measures what it purports to measure. The passages were selected with great care as representative of what pupils in high school read.

WHITE LATIN TEST

Another excellent test — the "White Latin Test"[1] — measures extent of Latin vocabulary and ability to translate Latin sentences. Its author is D. S. White. It comes in two duplicate forms, *A* and *B*. It is designed "to fill the need for a comprehensive and standard Latin test suitable to measure growth in knowledge of Latin on the part of high school and college students through four years of Latin and to provide a test which can be scored rapidly and accurately without the use of judgment on the part of the scorer."[2]

[1] Published by World Book Company (1924).
[2] *Manual of Directions*, White Latin Test.

The manual of directions accompanying the test provides the following description: "The test consists of two parts. Part I is a vocabulary test consisting of 100 words selected according to the frequency with which they occur in the works of authors that are read for college entrance. The frequency of occurrence of words in the vocabulary test is as follows:

First five words	1,000 times or more
Next twenty words	500 to 1,000 times
Next fifty words	100 to 500 times
Last twenty-five words	15 to 100 times

Part II is a test in the translation of sentences. Twenty sentences were composed and grouped according to increasing difficulty of syntax. The first ten should measure with considerable consistency ability of students of the first two years; the next five sentences, the ability of third-year students; and the last five sentences, the ability of fourth-year students. In a large measure the frequency with which these syntactical principles occur in high school Latin was the basis upon which the structure and sequence of the sentences were built."[1]

All items both in Part I and Part II are in the multiple-choice form, leading to perfectly objective scoring.

TABLE 46

WHITE LATIN TEST

SEMESTER NORMS

Semester	Part I	Part II	Total
1	29	30	59
2	39	45	84
3	48	57	105
4	56	65	121
5	63	72	135
6	70	77	147
7	76	81	157
8	82	85	167

[1] *Manual of Directions*, White Latin Test, p. 1. Copyright 1924 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

The score is the number right for Part I and five times the number right for Part II. Fifteen minutes are allowed for Part I and 20 minutes for Part II. On page 153 is a table of tentative norms obtained by testing 1,200 pupils who have had from one to eight semesters of Latin.

These norms show regular progression throughout the four years of high school. There is also plenty of room at the top and bottom to differentiate between poor students in the first year and good students in the last year of high school. Part I undoubtedly has high and satisfactory reliability, — the same cannot be said for Part II with only twenty items. The author has not supplied reliability figures.

These two tests measure the main direct outcomes of Latin instruction. There are, however, three other tests of special phases of Latin instruction of which mention must not be omitted, but which cannot receive extended discussion. The following brief descriptions give the salient facts about each test.

GODSEY DIAGNOSTIC LATIN COMPOSITION TEST

The "Godsey Diagnostic Latin Composition Test"[1] is composed of two parts: (1) thirty-three English sentences each followed by the Latin equivalent, one construction of which is repeated in four expressions, only one of which is correct; and (2) a set of rules for each sentence from which the pupil is to select the one applying to the form chosen as correct. In each case the pupil is to draw a circle around the form which he chooses as correct, and then around the number of the rule which he thinks explains the form chosen as correct; for example, sentence I–a reads:

The king's sons are leaders of the enemy.

Rēgis (fīlius, fīliōrum, fīliōs, fīliī) ducēs hostium sunt.

Rule 4. Possession may be expressed by the genitive.

Rule 7. The direct object is in the accusative.

[1] World Book Company, 1926.

Rule 13. The subject of a finite verb is in the nominative.
Rule 14. Action thought of as going on in the past is denoted by the imperfect.

The author is Edith Godsey. The test was first published in May, 1922, and Form II in April, 1923. The time allowance is 30 minutes.

PRESSEY TEST IN LATIN SYNTAX

The "Pressey Test in Latin Syntax"[1] was devised by L. C. Pressey of the Department of Psychology, Ohio State University. It is of the multiple-choice type, and consists of thirty-three English sentences, each of which is translated into four different Latin sentences, only one of which is correct. For example, Sentence 7, Form 1, reads:

I saw him in the city.
Eum in urbī vīdī. Eum urbe vīdī.
Eum in urbem vīdī. Eum in urbe vīdī.

The student is required to underscore the Latin sentence which he considers correct.

This test comes in only one form. Fifteen minutes is the time allowed.

TYLER-PRESSEY TEST IN LATIN VERB FORMS

The "Tyler-Pressey Test in Latin Verb Forms"[2] was constructed by Caroline Tyler and S. L. Pressey. It is of the multiple-choice type, and consists of thirty-two Latin verb forms. Each form is translated into four English verb phrases, only one of which is correct. The first form may illustrate the scheme.

1. Movēmus

We shall move. Let us move.
We are moving. We are moved.

It comes in only one form. It takes 15 minutes to give.

[1] Published by the Public School Publishing Company (1922).
[2] Published by the Public School Publishing Company (1922).

TABLE 47

GODSEY, PRESSEY SYNTAX, AND TYLER-PRESSEY LATIN TESTS: NORMS

	SEMESTERS STUDIED								TOTAL POSSIBLE SCORE
	1	2	3	4	5	6	7	8	
Godsey									
Sentences . .	7.8	14.2	15.8	18.7	21.3	24.6	24.0	26.3	33
Rules . . .	9.4	18.3	21.3	23.8	26.2	28.6	28.4	30.1	33
Pressey Syntax .	15.3	15.9	16.1	19.6	20.0	22.7	25.9	25.4	33
Tyler-Pressey									
Verb Forms .	13.7	17.5	19.8	19.2	22.0	23.8	62.6	25.2	32

Thanks to the Classical Investigation, Latin is supplied with well-constructed and serviceable standardized tests.

MODERN LANGUAGES

AMERICAN COUNCIL TESTS

The Modern Foreign Language Study has constructed standardized tests in French, German, Spanish, and Italian. These tests are known as the "American Council Tests."[1] I am indebted to Professor R. H. Fife for the following statement concerning the tests.

The Modern Foreign Language Study has prepared tests in two equivalent forms, *A* and *B*, each consisting of two parts (Part I, Vocabulary and Grammar; Part II, Silent Reading and Written Composition), for each of the three languages: French, Spanish, and German; and one form of two parts for Italian. The French, German, and Spanish vocabulary tests are based on word counts and contain a sampling of the most frequently used words in the language in ascending order of difficulty. The technique employed in the vocabulary tests for all four languages (French, German, Spanish, Italian) is designated as a "multiple-choice five-word technique." One or two of the five words bear a resemblance to the foreign word and are meant as "confusion words," calculated to mislead students unless they are sure of their ground. This technique is now widely used by educators since

[1] To be published by the World Book Company.

it makes it possible to test quite effectively the size of the student's vocabulary. Its advantages are:

a. The widest range of sampling (the greatest possible number of words) within the shortest possible time.
b. Unquestionable objectivity, *i.e.*, there is only one correct answer possible, hence the test can be scored easily even by persons unfamiliar with the foreign language, since a definite key can be constructed.

Objection has been made to this technique on the ground that a word may have many different meanings and its real significance can be ascertained only from a context, therefore, that the best way to test the knowledge of the meaning of a word is to have it appear, if not in a paragraph, then at least in a sentence. It has been shown, — on the strength of a wide testing with large numbers of students, — that there is a very close correlation between the scores of the sentence test and the isolated word test. Moreover, it is much more difficult, and frequently almost impossible, to construct an objective key with the paragraph or sentence technique.

The grammar tests are functional in character and make use of various test techniques. The material is representative of the grammar taught in high schools and in the first two college years. French employs largely the completion technique testing knowledge of functional grammar. Paragraph XII, however, is given to test the student's ability to recognize the correct form among four constructions, three of which are incorrect. In Spanish the recognition technique is used in Paragraph XIV, Form *A*, and in Paragraph IX of Form *B*. In German the recognition technique is used exclusively. Objections to the recognition technique are sometimes made on the ground that it is pedagogically unsound to place before students constructions which are fundamentally wrong. The reply is, that a test is not a teaching device but purely a means of sampling a student's knowledge. It has been shown that there is a close correlation between scores received by students on the completion and the recognition grammar tests. As to the possibility of objective scoring, the recognition technique is preferable because of its greater simplicity, students being required merely to underscore the correct construction or put a cross mark opposite it. The completion technique, on the other hand, necessitates the student's writing the answer. Frequently his handwriting is illegible, sometimes by intent, to conceal ignorance. This is especially true with the case endings of German pronouns, pronominal adjectives, and adjectives. The recognition technique does away with this objectionable feature.

The comprehension or silent reading test consists of paragraphs of increasing difficulty making use of the Thorndike-McCall paragraph reading technique. After some experimentation it was decided to call for answers in English, since in that way all doubt is removed as to whether the student really understands the question or merely picks out his answer mechanically from the paragraph.

On the last page of the comprehension test is a picture about which the pupils are requested to write a composition. Scales have been devised for grading these compositions: the French scale prepared by Professor Trabue of the University of North Carolina and the German scale by Miss Rossberg of Milwaukee Downer College under the direction of Professor Henmon. These scales are now ready for publication. The Spanish scales were to be ready in the autumn of 1926.

An examination of the norms and a study of the tests will convince one that modern language teachers now have a set of tools by which they can measure accurately the achievement of their pupils of three of the chief capacities involved in learning a modern language. The fact that students at different stages of advancement in high school and college can be tested by the same instruments and their scores directly compared is of the first importance for accurate classification on the basis of knowledge of the subject rather than purely on a time basis, as is too often our present practice. By comparison with the norms, teachers will be able to measure this achievement against national standards. Further, while the tests do not measure directly the student's ability to translate into English, this skill has been shown to correlate very closely with the results obtained from the vocabulary and grammar tests.

Norms are available for the French, German, and Spanish tests at the end of the semesters indicated.

The Modern Foreign Language Study is also issuing what they call "American Council *Beta* French and Spanish Tests."[1] These come in two forms. They were prepared by Jacob Greenberg and Ben D. Wood as tests to be used in a survey of the junior high schools of New York City. They are on a lower level of difficulty than the *Alpha* tests first described, and are to be preferred for testing first-year pupils. Each test has three parts. Part I is a vocabu-

[1] To be published by the World Book Company.

TABLE 48

AMERICAN COUNCIL MODERN LANGUAGE TESTS: NORMS

Test	Semester							
	1	2	3	4	5	6	7	8
French Vocabulary								
Q_3	20.7	27.4	33.3	37.2	44.1	47.4	51.9	60.8
Med.	16.2	21.7	27.4	31.4	37.3	41.1	43.1	52.3
Q_1	11.3	17.0	22.1	26.3	31.4	36.2	37.4	46.9
French Grammar								
Q_3	9.7	14.9	26.1	31.2	35.0	37.2	40.6	44.1
Med.	6.8	10.8	18.9	24.7	28.2	33.0	34.9	39.9
Q_1	4.5	7.3	12.3	17.2	20.7	26.5	27.4	33.4
French Reading								
Q_3	7.9	11.2	16.2	17.4	19.7	21.0	21.4	23.5
Med.	5.9	8.3	13.0	14.7	17.2	18.6	19.3	21.2
Q_1	4.0	6.1	9.4	11.8	14.4	16.2	17.0	19.0
German Vocabulary								
Med.	22.8	24.7	29.9	41.8	40.9	51.8		
German Grammar								
Med.	14.3	17.1	21.5	23.2	24.1	28.6		
German Reading								
Med.	8.4	11.9	12.9	17.1	19.1	23.9		
Spanish Vocabulary								
Q_3	30.8	34.7	44.0	50.5	58.5	59.3	66.3	73.9
Med.	23.4	28.2	37.0	43.6	49.7	49.4	59.9	65.8
Q_1	17.0	22.8	31.1	35.8	41.9	43.3	49.7	54.6
Spanish Grammar								
Q_3	13.6	17.2	22.9	25.1	30.1	29.7	38.1	40.1
Med.	9.9	13.1	16.3	18.4	24.1	24.2	31.4	33.0
Q_1	6.8	9.7	12.1	13.4	18.9	18.3	22.3	23.0
Spanish Reading								
Q_3	10.3	12.4	19.2	20.6	23.9	22.9	29.2	27.8
Med.	5.4	8.6	14.6	16.9	20.3	19.6	25.1	23.5
Q_1	3.3	5.7	10.7	12.8	15.8	15.7	21.3	19.6

lary test of 100 items built on the familiar multiple-response form. The selection of words is easier than in the Alpha American Council tests. Part II is a comprehension test of 60 items. One might describe this test as an information or common-sense multiple-choice test in French or Spanish. Examples of the items are:

French

a. Le chien est —
1. un livre, 2 un crayon, 3 un animal, 4 un ours, 5 un sac (3)

b. On se sert de l'encre pour —
1 manger, 2 boire, 3 courir, 4 nager, 5 écrire (5)

Spanish

a. El gato es —
1 un libro, 2 un árbol, 3 un animal, 4 una hoja, 5 una botella (3)

Part III is a grammar completion test of 60 or 65 items, consisting of short questions in English with answers in French or Spanish. These answers have a word deleted which is to be completed by inserting one, two, or three words in the blank. The deleted words have been chosen in order to test various grammatical principles. Examples of the exercises in Part III are:

French

a. I see some dogs
Je vois ( ) chiens (des)

b. I am hungry
( ) faim (J'ai)

c. I make him read this book
Je lui ( ) ce livre (fais lire)

Spanish

a. I am a man
( ) hombre (soy)

b. He is here
El esta ( ) (aqui)

Thirty minutes are allowed for each test.

The scoring of Parts I and II is perfectly objective, but knowl-

edge of the language is necessary to score Part III adequately. The length of these tests is such as to insure high reliability.

I should characterize the American Council Tests in modern languages as the most adequate tests available in any high school subject. They show what can be accomplished under coöperative effort.

REFERENCES

American Classical League — *The Classical Investigation, Part I*, Princeton University Press; 305 pp., 1924.

Brueckner, L., J. — "The Status of Certain Basic Latin Skills," *Journal of Educational Research*, 9 : 390–402, May, 1924.

Ullman, B. L., and Kirby, T. J. — "A Latin Comprehension Test," *Journal of Educational Research*, 10 : 308–317, Nov., 1924.

CHAPTER IX

STANDARDIZED TESTS FOR THE TEACHER OF SOCIAL SCIENCE

History

There are few standardized tests for the high school teacher of history. History seems to present stumbling blocks to the test maker. This lack of tests is partly due to the fact that there are no recognized definite aims or objectives for the study of history. The Committee on Social Studies in its report as one of the Commission on the Reorganization of Secondary Education in 1916 ignored the matter of objectives. Koos[1] has obtained from history teachers their conception of the aims of history teaching as given in the table on page 163.

The difficulty a test maker has is evident from the fact that the two aims most ignored by American History teachers are those of "mastering the text" and "equipping the student with a store of historical information," which are the two aims most easily tested. On the other hand, such aims as "promoting good citizenship" and "developing the faculty of discrimination," rated important by teachers, are impossible or difficult to test by objective methods. The history teacher objects to having a class held accountable for definite facts except the most general and obvious facts. There is inconsistency in this position for two reasons: first, history teachers do demand a mastery of the text and an acquaintance with historical information in the day-to-day recitation, yet they strenuously object to having a class tested on the results of this

[1] Koos, L. V. — "The Administration of Secondary School Units."

TABLE 49

PERCENTAGE OF CONCURRENCE OF TEACHERS OF HISTORY IN AIMS LISTED IN A QUESTIONNAIRE

AIMS	ANCIENT HISTORY	MEDIEVAL AND MODERN HISTORY	ENGLISH HISTORY	AMERICAN HISTORY
1. To master the text	59.2	44.2	52.9	46.2
2. To cultivate the power of handling facts	85.9	78.9	82.4	81.7
3. To develop the spirit of nationalism	42.3	40.4	47.1	70.2
4. To cultivate reconstructive imagination	59.2	55.8	82.4	59.6
5. To equip the student with a store of historical information . . .	62.0	65.4	64.7	57.7
6. To develop the "faculty of discrimination"	67.6	75.0	94.1	82.7
7. To promote good citizenship . .	73.2	69.2	70.6	93.3
8. To develop ability in speech, oral and written	76.1	65.4	58.8	76.0
9. To inspire with a love of reading	63.4	75.0	70.6	63.5
10. To teach the use of books . .	74.6	82.7	82.4	74.0

daily objective. Secondly, any of the wider powers of judgment — discrimination and reasoning — depend on a stock of valid yet specific facts, so that those despised facts are really fundamental to wider historical abilities. Tryon [1] says, "While the history teacher is waiting for the objectives in his subject to be scientifically determined, he will have to select from the unscientifically determined ones those he thinks of most value and see that he accomplishes them through his teaching." Unfortunately there is no scientific touchstone for the creation of educational objectives. These objectives depend upon individual wisdom.

[1] TRYON, R. M. — *The Teaching of History in Junior and Senior High Schools*, p. 216, 1921.

History teachers will never emerge from their confusion until they attempt to clarify their position.

Professor Henry Johnson has said some pertinent things relative to the type of question to be included in a history test.[1]

A history examination reduced wholly to memory questions is unreasonably unfair, and a standing indictment to reduce history teaching to memorizing. Facts are important. The American pupil should know more of them than he now seems to know, and know them more definitely. But the study of history, in any true sense, involves processes as well as a body of facts. To learn facts alone is in no real sense to learn history; to examine in facts alone is in no real sense to examine in history. . . . The memory test should not be allowed to dominate the history examination. It is, indeed, doubtful if more than a fourth, or perhaps a third, of the examinations should be devoted to tests of ability to *remember*. The remainder of the paper could then be devoted to tests of ability to *do;* to interpret a map or picture; to analyze a paragraph or a page of history; to find materials on a given topic; to solve by use of given materials a simple problem in criticism; to recognize in given facts differing degrees of probability; to judge from a given description some historical character; to discover in given conditions, past and present, resemblances, differences, relations, tendencies; to organize a given collection of facts; to select from the work of a term or a year, facts of special importance and to explain why they are important.

I am not holding a brief for the exclusive use of memory questions in history examinations. It may be that the objection to information questions is due to low pupil achievement in them. If that is the case, then history teachers ought to face the issue as other high school departments are doing. Low test scores would then be evidence that history teachers are trying to teach too many facts with a consequent failure. At least this is put forward as a plausible hypothesis.

There have been several attempts to standardize tests in history. Many of these tests are not commercially distributed, having never

[1] JOHNSON, H. — *Teaching of History in Elementary and Secondary Schools*, ch. 16, 1915.

emerged from their experimental status. The ones chosen for description here are available and have features of interest or value.

VAN WAGENEN AMERICAN HISTORY SCALES

The "Van Wagenen History Scales"[1] were constructed by M. J. Van Wagenen in 1919 as a doctor's dissertation at Teachers College. They still rank among the most serviceable of our standardized tests in American History. As originally published there were three series — information, thought, and character judgment. Each of these was prepared in two duplicate forms. (There was also a supplementary character judgment test.) In 1923 Van Wagenen issued a revised set of history scales which supersede the original scales. These revised scales are presented in as many as five duplicate forms in the information test series. Other tests were prepared dealing with limited periods of American History. The thought scales were untouched and the character judgment scales abolished. A new method of scoring was adopted (explained in detail below) which made these tests uniform with other tests developed by Van Wagenen.

They were originally constructed for use in the elementary school, but the revised scales include tests for the high school level of difficulty. Van Wagenen recommends using Information Scales R2, S2, T2, U2, or V2 for measuring a high school pupil's acquaintance with American History at the beginning of the year or during the early part of the year. A knowledge of the attainment at the beginning of the year would permit wise sectioning in case the class is to be divided. In any event, testing at the beginning of the year would tell the teacher much concerning individual differences. At the end of the year, Information Scales S3, T3, U3 should be used in the high school. The thought scales may also be used effectively in high school classes.

[1] Published by Bureau of Publications, Teachers College, Columbia University.

The original scales were largely of the question-and-answer type which necessitated an elaborate key. Part credits were originally given. A maximum credit of 3 was allowed on each question, and the key gave directions and sample answers for allotting credits of 1 or 2. When the scales were revised, past experience with using the scales was utilized, some questions were omitted and others altered so as to make the scoring key less confusing and less difficult. For the revised scale a maximum credit of 2 is allowed and the key gives directions for determining answers worthy of a credit of 1. In order to prepare the additional forms, many more tasks than were in the original scales were assembled, tried out, and selected "not only from every period of American History, but also from every phase or point of view."

Samples of the questions in the information scales are:

1. From what country was each of these territories acquired by the United States?
   a. Alaska
   b. Louisiana
   c. Porto Rico
   d. Florida
2. Of these conditions, put a check mark in front of those which led to the financial panic of 1837.
   —— Growing population
   —— Speculation in land
   —— Introduction of the "spoils system"
   —— "Issuing of Species" circular by President Jackson

Samples of questions in the thought scales are:

1. It was not until about 1887 that electric power began to come into use. After 1887 there was a great increase in street railway building. In the cities of Massachusetts there were three times as many miles of street railways in 1897 as in 1887.
   What was evidently one of the things to which this sudden rapid increase in street railway building was due?
2. We know that in 1790 there were about 2,929,200 people living in the

United States. We do not know at all accurately how many people there were in the colonies at any previous year after 1650.

Why do we know how many people there were in the United States in 1790, but do not know how many there were in any year before that?

We are told that the difficulty values of the separate tests were found by testing about 500 pupils in the upper three grades of a typical Minneapolis public school, a typical Kansas City public school, and about 150 other high school students, in American History. About 140,000 answers thus obtained were scored.

Forty minutes are allowed for each scale. Van Wagenen recommends taking a small number of tasks, say three or five, and correcting these on all of the papers, when correcting a set of papers.

The scoring of the scales is peculiar. It is a system devised by Kelley [1] to determine by interpolation the point on a scale where an individual can do 50% of the items of that difficulty or quality. Van Wagenen later greatly simplified this method and used it in his tests.

Each scale is divided into three groups of tasks called Group I, Group II, and Group III, which represent three levels of difficulty. Each group or level is given a difficulty value. A difficulty value of 75 corresponds to the average achievement of 12-year-old children. Ten units on this scale is the equivalent of one quarter of the distribution of 12-year-olds. The score a pupil finally receives comes out in terms of this unit.

There are ten tasks in each group. In finding a pupil's score one finds the errors or omissions made in each of the groups. By referring to a table one finds the score corresponding to the number of errors in the hardest group. From this score is subtracted the number of errors made in the other two groups. This is called the pupil's uncorrected score. If the pupil makes

[1] Kelley, T. L. — "A Simplified Method of Using Scaled Data for Purposes of Testing," *School and Society*, 4: 34–37, 71–75, July 18, 1916.

errors in Group I, it is probable that he would make errors in a group of difficulty level lower than Group I. Accordingly, a correction must be made when there are errors in Group I by referring to a second table and subtracting a value found in that table, from the first uncorrected score.

TABLE 50

VAN WAGENEN AMERICAN HISTORY SCALE: "SCORING TABLE"

When errors in Group III are .	½	1	1½	2	2½	3	3½	4	4½	5	5½	6	6½	7	7½	8	8½	9	9½	10
Pupil's uncorrected score is . . .	113	108	104½	101½	99½	98	96½	95	93½	92	91	90	89	88	87	86	85½	84½	83½	83
When errors in Group I are . .	½	1	1½	2	2½	3	3½	4	4½	5	5½	6	6½	7	7½	8	8½	9	9½	
Take from First Corrected Score .	0	0	½	1	1	1½	2	2½	3	3½	4	5	6	7	8½	10	12½	15	20	

Suppose, for example, that the errors a pupil makes are as follows:

Group I	4
Group II	6
Group III	8

Step I — By referring to Table 50 one finds that the score corresponding to 8 errors in Group III is 86.

Step II — The sum of errors made in Groups I and II is 10, and 10 subtracted from 86 is 76.

Step III — By referring to Table 50 one finds that for 4 errors in Group I 2½ points must be subtracted from this first corrected score. 76 — 2½ is 73½. This is the pupil's score.

Van Wagenen claims that his system of scoring has marked superiority over the common method of simply totaling correct responses. For one thing, "differences in one part of the scale are directly comparable with differences in any other part of the scale."[1] This factor, so important in a true scale, loses its potency in educational measurements where an individual score has such a high probable error. The other advantage is that

[1] VAN WAGENEN, M. J. — "Some Implications of the Revised Van Wagenen History Scales," *Teachers College Record,* 27 : 142–148, Oct., 1925.

the range of difficulty is narrow. There are no very simple items that are seldom missed, nor any very difficult items that are seldom done. This is a true advantage, but one that is dependent on the selection of items and not on the method of scoring. Any test may have the same advantage that has the same distribution of difficulty of the items.

Norms are available from the fifth through the eighth grades.

Unlike many standardized history tests, the Van Wagenen scales represent good or acceptable history. Van Wagenen was careful to have his items gone over and criticised by history experts before inserting them in his test. The tests can be criticised from the test maker's point of view for not being as objective as possible. A more extended use of the multiple response type of question would make the scoring more objective and easier, without detracting from the value of the test.

BARR DIAGNOSTIC TESTS IN AMERICAN HISTORY

An attempt to get away from the purely informational type of test is represented in the "Diagnostic Tests in American History"[1] constructed by A. S. Barr. Although this test has not been widely used, it possesses novel and suggestive features that make it worth describing. The first edition was printed in 1920 and it has now gone through three revisions. It comes in two forms, *A* and *B*, similar in construction and arrangement and of equal difficulty.

There are five tests: (1) comprehension, (2) chronological, judgment, (3) historical evidence, (4) evaluation of facts, and (5) causal relationships. The author makes the following statement of the reasoning leading up to his choice of tests.

When the history pupil is confronted with the printed page, he must first read and understand the material placed before him. History is a reading subject, but differs from general reading in that it employs historical back-

[1] Published by Public School Publishing Company.

grounds and a technical language. Then, secondly, the really "historically minded" person asks: "Are the facts here stated true?" Pupils frequently pass over this step with a vague general feeling that all printed facts are true. The third inquiry is: "Is the fact important?" Even the most immature pupil will make some selection. Few choose to memorize the text "verbatim." The fourth step is to place the facts in time order. The facts of history are peculiar in that they are all placed in a definite time relationship. When taken out of this time relationship, they lose their historical significance. And finally the facts of history must be studied in their relation to other facts. The facts of history are not isolated facts, but rather trains of facts tied in causal relations. Five tests are given — each to measure one of these five fundamental processes.

That the reader may see that the tests are unique, samples of each are given:

Test I

1. Read through the following paragraph rapidly. Answer as many as you can of the questions which follow. Read the paragraph again if you need to.

The Yorktown campaign of 1781 practically marked the end of the Revolutionary war. General Clinton with the main British army was stationed at New York. General Cornwallis, after chasing Lafayette from one end of Virginia to the other, returned to the seaboard and fortified himself at Yorktown. De Grasse, the commander of the French fleet, was in the West Indies. Word came to Washington, who was keeping watch over Clinton at New York, that De Grasse would come north and be at the mouth of the Chesapeake on the first day of September. Washington, although he would have preferred to attack Clinton at New York, left a guard at West Point and moved southward to the head of Chesapeake Bay. The march was so skilfully executed that Clinton believed an attack was contemplated upon his army at New York. The three allied forces under Washington, Lafayette, and De Grasse concentrated their attack upon Yorktown. Cornwallis, besieged from both land and sea, was forced to surrender in October, 1781. The surrender of Cornwallis practically ended the war.

(*a*) Put a cross (×) before each of the following statements that is discussed in the above paragraph.

(1) The French and Indian wars
(2) The siege of Yorktown
(3) The surrender of Cornwallis

(*b*) From the following list of names, draw a line under those that are names of British generals.

WASHINGTON LAFAYETTE
CORNWALLIS DE GRASSE
CLINTON

(*c*) Draw a line under one of the following words which tells who lost in the engagement at Yorktown.

FRENCH ENGLISH AMERICANS

(*d*) Draw a line under the name of one of the following cities which tells where Clinton's army was stationed.

YORKTOWN NEW YORK
PHILADELPHIA

(*e*) Draw a line under the most important event in the above paragraph.

Test II

1. Indicate the chronological order (time order) in which the following men lived by placing a figure one (1) before the individual who lived first, a figure two (2) before the individual who lived second, and so on through the list.

(*a*) Patrick Henry
(*b*) Robert E. Lee
(*c*) John Cabot
(*d*) James Monroe
(*e*) Roger Williams

Test III

1. Read through the reading passages hurriedly. Concentrate upon the questions and directions.

If you were writing an account of the burning of a county court house in 1880 and came to a point in the story where there were the three following conflicting accounts, which of the accounts (from all you can learn from the accounts themselves) would you accept as probably the correct version of the story?

Put a figure one (1) before the account which you think is most likely to be the correct one; a figure two (2) before the account that is next in value, and a figure three (3) before the one that is the least likely to be the correct version.

(*a*) Mr. Jackson, an old resident of the county-seat who was an eye witness of the affair, said that the building burned rapidly and that but few of the records were saved.

(*b*) An old letter written at the time by Mr. Hough to a friend, and an account in the local paper, say that all the records were saved.

(*c*) Mr. Smith, another old resident, said that in his opinion none of the records were saved.

Test IV

1. Put a cross (×) before the event in the following list which has been of the greatest importance in American History.

(*a*) The founding of Philadelphia
(*b*) The purchase of Alaska
(*c*) Jay's treaty with England
(*d*) The Declaration of Independence
(*e*) The election of Andrew Jackson

Test V

1. The United States was at war in 1863, 1898, and 1917. In 1840, 90,000 immigrants came to America; in 1855, 430,000 immigrants came; in 1863, 100,000 immigrants came; in 1867, 350,000; in 1882, 788,000 came; in 1898, 225,000 came; in 1913, 1,200,000; in 1917, 200,000.

Put a cross (×) before two of the following conclusions that might be drawn from the above facts.

(*a*) Immigration into the United States increases during periods of war.

(*b*) There has been a general increase in the immigration into the United States during the period 1840–1917.

(*c*) War reduces immigration.

(*d*) Immigration into the United States decreased during the period 1840–1917.

(*e*) The greater majority of the immigrants to the United States since 1900 came from southern and eastern Europe.

The test has not been standardized. The difficulty value of the items is not known. There is a practice exercise for which five minutes is allowed. This is not scored and the pupils are told of this fact in giving the test. Six minutes is allowed for each of the five tests. A rather complicated system of scoring by weighting the exercises is used. Apparently the weights have been determined in an a priori manner.

Norms are furnished for senior high school history, but one is not informed as to the number or sampling in determining the norms.

TABLE 51

BARR DIAGNOSTIC TESTS IN AMERICAN HISTORY: NORMS

Test	1	2	3	4	5
Scale 2B	10.0	8.0	9.0	12.3	7.5
Scale 2A	8.2	7.5	6.2	12.2	8.9

This test can be praised for giving good suggestions for the construction of tests in history, but it cannot be recommended as a standardized test. As a standardized test it is not finished. In particular we are not told how reliable it is, nor how well it correlates with other measures of history ability. The probabilities are that the reliability is not high and that it does not justify its title of diagnostic.

PRESSEY-RICHARDS TEST OF UNDERSTANDING OF AMERICAN HISTORY

Another suggestive test is the "Test of Understanding of American History"[1] devised by S. L. Pressey and R. C. Richards. It comes in one form only.

[1] Published by the Public School Publishing Company, 1922.

It consists of four tests: Test 1, character judgment; Test 2, historical vocabulary; Test 3, sequence of events; Test 4, cause and effect relationships. Following are samples of each:

Test 1. Character judgment.

1. Columbus: cowardly proud <u>courageous</u> rich

Test 2. Historical vocabulary.

1. What is a stockade?

   A part of a gun;
   A supply of ammunition;
   A public punishment;
   <u>A barrier for defense.</u>

Test 3. Sequence of Events.

(In this test the pupil is to underline the event that happened the longest time ago.

1. Discovery of the Pacific; Discovery of the Mississippi; <u>Discovery of America</u>; Discovery of Hudson Bay.

Test 4. Cause and Effect Relationships.

(In this test there are the names of three events that acted as "causes" and one event which was the "effect" of the operations of the three "causes." The effect is to be underlined.)

1. Religious persecution <u>Colonization of America</u> Desire for adventure Desire for political independence

Each test contains 26 items. The first item in each test is used as a sample or illustrative exercise, leaving 25 items to be scored. This makes 100 scorable items for the whole test. The time allowances are given in the next table.

Concerning the interpretation of scores the authors say:

Poor work on the first test indicates a lack of appreciation of the character of the great men in American History and such appreciation is usually considered one of the objectives in the teaching of history. Low scores on Test 2 would suggest such a small technical vocabulary of historical terms as seriously to interfere with the understanding of the textbook; intensive drill in the meaning of unusual words should be tried. Low scores on the third test would indicate an insufficient drill on the basic facts of American

History — all the facts used appear in several texts used as a basis for the test. Poor work on Test 4 suggests a failure to emphasize the relationships of historical events.

These interpretations are only guesswork and indicate a prevailing tendency to jump at conclusions with regard to the results of history teaching. No one knows, for instance, whether inability to judge correctly the character of historical personages indicates a lack of appreciation of the characters of these men or whether what the test measures is an important objective of history teaching. Such matters are for research to decide.

TABLE 52

PRESSEY-RICHARDS TEST IN AMERICAN HISTORY: TESTS, TIME ALLOWANCE, AND OCTOBER NORMS

Norms are given for the senior class in high school in October:

TEST	TIME ALLOWANCE IN MINUTES	NORMS FOR 12TH GRADE (OCTOBER)
1	5	15
2	6	17
3	6	16
4	8	15
Total	25	63

The test should be criticised as being one of a number of tests thrown together with no explanation of the method of assembling the material, or of the order of items; and no statement as to reliability or validity. In this particular instance it would seem that it should have been easy to assemble enough material to make duplicate forms. Notwithstanding our ignorance of these matters, the test appears to be useful. The multiple-choice form makes it objective and the large number of items would tend to make it fairly reliable. The tests themselves are suggestive and would seem to measure important phases of history teaching, but these are only statements and the facts need to be determined experi-

mentally. All in all, this test seems to be worth considering. In a recent study Ruch finds it to be the best of the standardized tests in American History.

KEPNER BACKGROUND TESTS IN SOCIAL SCIENCE

This test[1] is not solely, but largely, a history test. The author states:

> The tests aim to measure the extent of the factual background in the social sciences which a pupil or group of pupils bring to *any* one of the social science subjects in the secondary school. The tests would aim at determining the social science background of the pupils by measurement rather than follow the usual practice of assuming an empirical background. The logical and pedagogical time for such an inventory is at the beginning of the course. The content of the present tests represents an effort at selecting fair samplings of a practical background and comprises material from the fields of history (especially American), government or civics, economics, geography and sociology, with emphasis roughly in the order of the subjects named. The core of the tests is history.

The author is Tyler Kepner. There are two forms, *A* and *B*, similar in construction and content. The tests are useful before beginning the course in American History.

There are seven groups of questions called exercises. Exercise I tests the association of men-events and events-men. The multiple-choice form is used. Exercise II tests literary background. Here again the multiple-choice form is used. Exercise III tests geographic concepts. The completion form is used. Exercise IV tests knowledge of historical words and terms commonly used. The best answer form is used. Exercise V tests knowledge of civic, social, and economic terms. The matching type of test is used. Exercises VI and VII test dates and chronology and the test form is one in which items have to be placed in order. An ordinary school period of 40 minutes is allowed. In scoring, the number of correct responses is the unit

[1] Harvard University Press, Cambridge, Mass., 1924.

used. Each item counts one and no part credits are allowed. (This applies to sections 6 and 7.) Norms are not given in the manual. The maximum score is 69.

Kepner[1] tells us that the items have been very carefully selected. Twenty-one representative text-books were tapped for the source of the questions. (Just what subjects the texts represent we are not told.) The original questions, numbering 200, were reduced to 176 by the critical analysis of teachers. These items were administered in eleven schools to 580 ninth-grade pupils and 220 eleventh- and twelfth-grade pupils. The items in each exercise are arranged in order of difficulty.

The author does not state the reliability of his test. Because of its length it probably contains fair reliability. Little or no dependence can be placed on the score in any one section. Evidently the author intends that the score on each section shall be used. He says: "One word of suggestion may be added regarding the interpretation of individual scores by teachers. Attention should be given not only to the total score, but also to the scores of the several exercises."[2] It may be categorically stated that the separate exercises provide too narrow a sampling to be individually diagnostic. If the purpose of the Kepner test is diagnostic, the test should stick to a single type of exercise. It is well known that the multiple-choice and best-answer type questions differ in difficulty, and in varying the type one is sure to get a wrong impression as to the proficiency in one or another of the exercises. For diagnostic purposes the single-answer or best-answer type of test is to be preferred.

One of the difficulties of making a standardized test in history is well illustrated in this test. Item 6 of Exercise I, form B, reads:

6. Geographical discoveries in Columbus' time were made easier by the invention of

Paper Mariner's Compass Steamboat Searchlight

[1] Cf. *Manual of Directions.* [2] *Ibid.*

Mariner's compass is the answer intended, but *paper* is also a true answer. Indeed, although the compass is perhaps more immediate it would be impossible to say that *compass* is more true than *paper*.

Or take item 9 in the same exercise:

9. The first man to sail around the world was
   Columbus Balboa DeSoto Magellan

The answer intended here is *Magellan*, although it is a historical fact that Magellan died and was buried in the Philippines, and not he, but his expedition, finally encircled the globe. In school texts Magellan is usually credited with being the first to sail around the world. Hence a boy will be denied credit if he fails to give the conventional answer and shows a knowledge of history on a higher level. Historians in general criticize standardized tests in that the tests penalize too great a knowledge of history and encourage conventionality. The tests do not provide for free discussion and the exhibition of initiative.

This discussion represents the chasm between the psychologist's point of view and the subject-matter specialist's point of view in measuring achievement. The historian is anxious to learn whether or not his pupils can think historically. The psychologist is interested in sorting out a group of students according to historical ability. He depends on his correlations. If the correlations are high between his test results and the order in which the teacher will place the class, he is satisfied. The historian cares little for these statistical criteria and judges a test entirely by its historical value. Take, for example, the two questions cited above. They are valid questions for the psychologist if they tend to sort out the better history students from the poorer. It goes back eventually to the purpose of testing. If the purpose is to find out how much the pupils know or can do, then the subject-matter specialist's type of test is best. But if the purpose is to mark, section, or

advise pupils, then the psychologist's type of test is best. The historian's purpose leads to the diagnostic test — the psychologist's purpose leads to the achievement test. Each is legitimate in its sphere.

All of the tests so far described have been in American History. A good test in ancient history has been constructed by Hackett, but it is not at present in print for commercial distribution.

Vannest has constructed a diagnostic test in modern European history. The test will not be described in detail.

Civics

A subject like civics presents unusual difficulties to the test maker. In looking over a civics test with a view toward selecting test elements, one is surprised at how really few significant statements there are. Either the facts are detailed local or statistical facts that no one is expected to remember or there are vague statements about trends or tendencies. There are few direct statements that could be made a basis for testing.

BROWN-WOODY CIVICS TEST

The "Brown-Woody Civics Test"[1] represents a careful piece of work in constructing a standardized test in civics. The authors are Arnold W. Brown and Clifford Woody. There are two forms, *A* and *B*.

The Civics Test is based primarily upon the subject matter which is emphasized in the teaching of civics in the high school. Each exercise of the test is founded upon subject matter common to at least five of nine of the most widely used textbooks in civics, which were carefully and minutely analyzed preliminary to the construction of the test. In the construction of the different tests there was a definite attempt to select exercises emphasizing understanding of the vocabulary of civics, mastery of the facts of civics, and the application of the principles of civics to practical situations. Stated in slightly different fashion, the exercises emphasize information,

[1] Published by the World Book Company (1926).

habits, thinking, ideals, attitudes, and appreciations. Although only three tests have been constructed, it is evident that all these aspects of civics are included.

PART I. *Civic Vocabulary.* Test I is designed to determine the extent of the pupil's mastery of the vocabulary encountered in the literature of civics. It consists of forty words selected at random from the list of words common to at least five of the nine textbooks, which were considered essential to an intelligent reading of the subject matter of the civics textbooks. This is an "alternative answer" test, in which the pupil must select one of four given words or expressions which most nearly means the same as the indicated word in the vocabulary test. In building this test there was a definite attempt to incorporate words or expressions which were closely enough related to the given word in meaning or in form to insure a careful discrimination in choosing the proper response.

PART II. *Civic Information.* Test II is designed to measure the extent of the pupil's information in the special field of civics. It consists of eighty "yes or no" questions common to at least five of the nine textbooks. The questions are distributed over local, state, and federal government in approximately the same proportion as revealed in the analysis of subject matter in the textbooks. By including as many as eighty questions this test is of sufficient length to give reliable measurement.

PART III. *Civic Thinking.* Test III is designed to measure how well the pupil can apply the principles of civics to practical situations. The responses to these situations demand the application of what may be termed "civic thinking" and at the same time it is possible through them to determine whether the pupil possess proper civic attitudes, ideals, and appreciations.[1]

The most questionable section of the test is Part III. One novel item tells, by means of descriptive adjectives, of two candidates running for the office of mayor. The pupil is told to place an × in the square following the name of the candidate for whom he would vote, also to check the descriptive items which determined his choice. No use is made in the scoring of these checks. The rest of Part III puts such a premium on reading

[1] *Manual of Directions*, Brown-Woody Civics Test. Copyright 1926 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

ability that it is probable that the real purpose of the test is submerged.

An over-all time of 35 minutes is allowed. The score on Parts I and II is the number of correct responses. The score on Part III is found by multiplying the number of correct responses by 3. The maximum scores are, Part I, 40; Part II, 80; Part III, 27.

TABLE 53

BROWN-WOODY CIVICS TEST: PERCENTILE NORMS

(End of Year)

Percentile Rank	Senior High School	Junior High School
90%	132	118
75	126	110
50	118	99
25	108	88
10	99	77

This chapter has been filled with the criticisms made of standardized tests for the measurement of history. Let us close by briefly summarizing these criticisms.

1. *The present standardized tests in history measure only memory. The questions are only information questions.* A study of the Barr test and the Pressey-Richards test shows that the objective test can be made to measure other phases of historical ability.

2. *Many of the facts called for or emphasized are not always of the most importance.* This, where true, is a serious indictment against standardized tests. When true, it is a sign that the tests have been hurriedly and carelessly assembled.

3. *Tests are too long for the allotted time.* This criticism shows an ignorance of the technique of testing by the critic.

4. *The tests can be used only once.* This is not an exclusive feature of standardized tests. Of course our position is that no

standardized test has been perfected until there is a duplicate form.

5. *Difficulty is encountered in the scoring of standardized tests.* The retort is that if difficulty is encountered, it is a lesser difficulty than that encountered in grading the traditional test.

6. *Standardized tests usually cover a whole historical division,* i.e., *American History. Such a test cannot be used with profit until the end of the year.* This criticism discloses another prejudice with regard to testing. The point is that teachers do not like to acknowledge that pupils may know more history than has already been covered to date in the class. Is there a penalty against reading ahead in the text? Is there any spur to reading ahead in the text? From one point of view tests which cover a whole course are good at any time in the course if only to spur on the class to a maximum achievement. It must be admitted that tests can be made more reliable if they are made to fit the subject being tested. On the other hand, there is a decided advantage in making achievement seem a progressive or growing affair, as one can by using a test at any time covering the work of the whole year.

7. *The teacher stresses the material of the test in teaching.* This criticism cannot be confined to standardized tests.

8. *Guesswork creeps into many of the new-type tests.* Here again the new-type tests are not the only offenders.

REFERENCES

Barr's "Diagnostic Tests in American History," *Journal of Educational Research*, 3: 71–75, Jan., 1922.

Gibson, O. H. — "Existing Standard Tests in History," *Historical Outlook*, 12 : 324–326, Dec., 1921.

Hardy, R. E. — "New Types of Tests in Social Science," *Historical Outlook*, 14 : 326–328, Nov., 1923.

Johnson, H. — *Teaching of History in Elementary and Secondary Schools*, ch. 16. The Macmillan Company, New York; 497 pp., 1915.

KELLEY, T. L. — "A Simplified Method of Using Scaled Data for Purposes of Testing," *School and Society*, 4 : 34–37, 71–75, July 1, 8, 1916.

— "Individual Testing with Completion-Test Exercises," *Teachers College Record*, 18 : 371–382, Sept., 1917.

KEPNER, P. T. — "A Survey of the Test Movement in History," *Journal of Educational Research*, 7 : 309–325, April, 1923.

KOOS, L. V. — "The Administration of Secondary School Units," SUPPLEMENTARY EDUCATIONAL MONOGRAPHS, VOL. 1, NO. 3, University of Chicago, July, 1917.

RUCH, G. M., and others. — *Objective Examination Methods in the Social Studies*. Scott, Foresman and Company, Chicago; 116 pp., 1926.

RUGG, E. U. — "Character and Value of Standardized Tests in History," *School Review*, 27 : 757–771, Dec., 1919.

Social Studies in Secondary Education — *Report of the Committee on Social Studies of the Commission of the Reorganization of Secondary Education of the National Education Association*, U. S. BUREAU OF EDUCATION BULLETIN NO. 28, 1916.

VAN WAGENEN, M. J. — *Historical Information and Judgment in Pupils of Elementary Schools*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION NO. 101, Bureau of Publications, Teachers College, Columbia University, New York; 74 pp., 1919.

— "Tables for Computing Mean Individual Scores in Educational Scales," *Teachers College Record*, 21 : 441–451, Nov., 1920.

— "Van Wagenen American History Scales, Manual of Directions," Revised Edition, TEACHERS COLLEGE BULLETIN NO. 4, FIFTEENTH SERIES, Oct. 20, 1923.

— "Some Implications of the Revised Van Wagenen History Scales," *Teachers College Record*, 27 : 142–148, Oct., 1925.

VANNEST, C. G. — "Diagnostic Test in Modern European History in Eighth Annual Conference on Educational Measurements," Indiana University, Bloomington. *Bulletin of the Extension Division, Indiana University, Vol. 6, no. 12*, pp. 59–68, 1921.

CHAPTER X

TESTS FOR PHYSICAL EDUCATION

Objectives

In no department of secondary education is there more confusion as to aims than in physical education. The aims as stated try to cover too much and are vague and indefinite. Williams [1] tells us that "Physical education should aim to provide an opportunity for the individual to act in situations that are physically wholesome, mentally stimulating and satisfying, and socially sound." This aim gives no conception of the activity physical education consists of; it is so broad that it might fit many different subjects.

It is quite evident that those interested in physical education are aiming at something beyond the mere training given during the physical education period, but they do not state these aims in definite terms. In order to make clear the confusion that exists regarding the adjectives of physical education I quote the following statement from Bobbitt: [2]

> The physical character of our population proves that physical education is not yet effectively performing its function. Its aims have been too vague. Its program has been too incomplete. Physical training departments prefer to take care of a program of gymnastics, games, sports, athletics, etc., which constitute only a fraction of the total program; and even here they are coming to place foremost not the physical objectives but the social objectives. This is well indicated by the recent report of the Committee of the Society

[1] Williams, J. F. — *The Organization and Administration of Physical Education*, p. 18, 1923.

[2] Bobbitt, F. — "Objectives of Physical Education," *American Physical Education Review*, 26: 229–231, May, 1921.

of Directors of Physical Education in Colleges on the aims and scope of physical education.

"Aims. 1. If we conceive the perfecting of the individual in his social relations to be of greater importance than more purely personal values, we may well begin our list of aims with certain qualities developed by appropriate group activities, particularly games and athletic sports, practiced under favorable conditions. It is through these agencies that the child and youth most readily and naturally acquire habits of obedience, subordination, self-sacrifice, coöperation and friendliness, loyalty, capacity for leadership, ability to lose without sulking and win without boasting, a spirit of fair play, and all that is implied in the word sportsmanship.

"2. Other qualities of marked, though indirect, significance to the community are self-confidence and self-control, mental and moral poise, good spirits, alertness, resourcefulness, decision and perseverance, courage, aggressiveness, initiative. These traits, developed by the farm life and varied home activities of an earlier age, must now be insured through other means than those which the average family can itself supply.

"3. Underlying such aims must be the purpose to promote the normal growth and organic development of the individual, to conserve his health and provide a fair degree of strength and endurance, and to secure an erect and self-respecting carriage of the body and the neuro-muscular control required for prompt and accurate response and graceful and effective movements. Emergencies should be anticipated by training in exercises of which swimming may be taken as a type, and by others which accustom one to bear physical punishment coolly and to defend oneself successfully."

Here the leaders of physical education in this country present their most frequent formulation of their objectives under three heads; and two of these are social and psychological and not physical at all. And what is more, these are placed first, the actual physical objectives trailing the procession rather lamely.

There appears to be a feeling among physical educationists that the physical side of man's nature is lower than the social and the mental; and that if they would raise their department in the eyes of all concerned, they too must aim primarily at those more exalted, non-physical things of mental and social type. Apparently something needs to be done by way of bringing physical educationists to value more highly the physical side of men.

Since there is confusion in the minds of the physical educationists as to what they are trying to accomplish, one would expect to find a similar confusion in the measurements of the results of physical education, and this is actually the case. Some speak of the estimation of health, others of tests of physical efficiency, others of measures of motor ability, others of athletic ability, still others of physical condition. In each case the measures are of conveniently measurable things, but no one has taken the trouble to make sure that what is being measured has any relation to educational objectives. Indeed, I find some physicians and physical education directors remarkably obtuse to any conception of scientific measurement of the products of education.

Physicians especially resent the interference of outsiders in what they consider their technique of diagnosis. The medical profession seems to be steeped in tradition and averse to having its diagnostic methods validated by methods used in other scientific fields. Physicians resent the intrusion of outsiders into their field and yet many themselves lack the technique of statistical research. Talk with any physician about this and he will digress to a discussion of the individual exceptions to any statistical law derived from measurements. Because measurements possible at the present time are only first approximations, the correlations are low and consequently show trends only, but at least these trends are beyond cavil. Physical education directors as a class are more open-minded, but their training has been wholly at the hands of the physicians.

Of all the school records which appear useless the records in the office of a physical education department in a high school seem most useless. Naturally many hours have been spent by a physician and the physical education director in obtaining anthropometric measurements — heights, weights, girths, capacities, etc. — a battery of strength records and diagnoses of the conditions of the sight, hearing, throat, lungs, heart, skin, glands, etc. There the

records lie, testimony to the industry of those who accumulate them.

Out of the welter of measurements in the physical education field, the following types may be classified:

a. Physiological measurement.
b. Anthropometric measurement.
c. Strength measurement.
d. Athletic proficiency measurement.

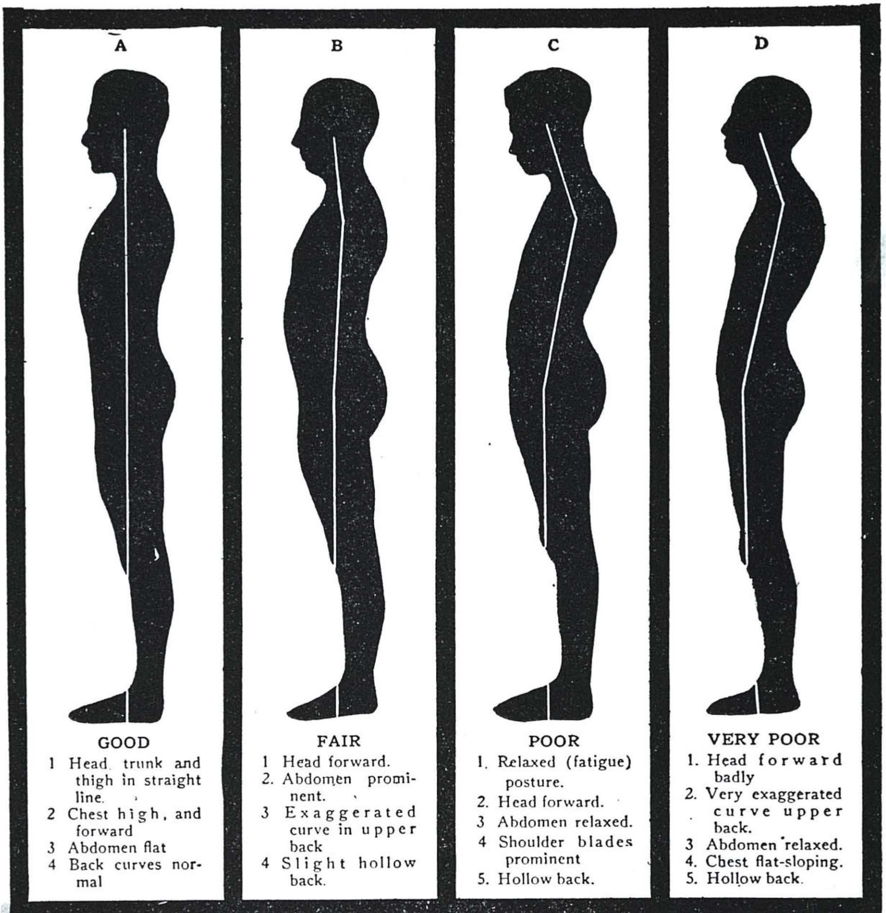
Just how these are related to the objectives of physical education in the high school is impossible to state with our present knowledge and with the chaotic condition of objectives of physical education.

Physiological Measurement

A physician's diagnosis of the condition of the eyes, ears, mouth, lungs, feet, and spine probably constitutes a valuable index of general health. In general a physician's estimates are descriptive. There are, however, devices for expressing the strength of vision quantitatively. There has been worked out a scale for the measurement of posture. Following are photographs designed to serve as a scale for the measurement of posture.

Rating methods can easily be worked out for most of a physician's diagnoses, but the scales for rating need to be carefully standardized.

Perhaps the most important of these physiological measurements are those connected with the circulatory system. One of the first of these is the "Foster Test of Physical Efficiency," based on certain features of pulse rate. Foster gives standards which enable one to determine good, average, and bad physical efficiency. His test fails in usefulness because he has given only the vaguest means of interpreting the results of testing any one boy.



Courtesy of the Department of Education, University of Southern California

FIGURE 1. — Posture Scale

THE SCHNEIDER INDEX

By far the most successful attempt to measure the cardio-vascular system is the point measure of E. C. Schneider, known as the "Schneider Index." This index was the result of research at the Medical Research Laboratory, Air Service, Mitchel Field, Long Island, in an endeavor to find an easy method of assessing the physical fitness of aviators.

The procedure is as follows:

1. The patient reclines for five minutes. (*a*) The heart rate is then counted for twenty seconds. When two consecutive twenty second counts are the same, this is multiplied by 3 and recorded. The score is noted according to Part A, Table 54. (*b*) The systolic blood pressure is next taken by auscultation; two or three readings are made as a check.

2. (*a*) The patient stands at ease for one or two minutes to allow the pulse to assume a uniform rate. When two consecutive twenty second counts are the same, this is multiplied by 3 and recorded. The score is obtained by use of Part C, Table 54. The difference between the standing and reclining pulse rates is scored then by use of Part B, Table 54. (*b*) The standing systolic pressure is next taken. The difference between this and the reclining systolic pressure is then scored by Part F, Table 54.

3. The patient next steps on a chair about 18 inches high, five times in fifteen seconds timed by a watch. To make this test uniform, he stands with one foot on the chair at the count one; this foot remains on the chair and is not brought to the floor again until after the count five. At each count he brings the other foot on the chair and at the count "down" replaces it on the floor. This should be timed accurately so that at the fifteen-second mark both feet are on the floor. (*a*) Immediately while he stands at ease, the pulse rate is counted for fifteen seconds; this is multiplied by 4 and recorded. (*b*) Counting is continued in fifteen-second intervals for two minutes, record being made of the counts at 60, 90, and 120 seconds.

The data from (*a*) will be scored by Part D, Table 54, taking the difference between this exercise pulse rate and the standing rate. The data in (*b*) are scored according to Part E, Table 54.

Schneider's index is inadequate because he has provided only partial means of interpreting the results. By measuring 54 men with the index who were reported by physicians to be ailing and below standard he found that approximately 89% made scores of 9 or below. Accordingly he makes 9 the critical score. "A score of 9 or less gives indication that the clinician may find something wrong with the patient."[1] This arbitrary point is better than

[1] SCHNEIDER, E. C. — "A Cardiovascular Rating as a Measure of Physical Fatigue and Efficiency," *Journal of American Medical Association*, 74: 1507–1510. May 29, 1920.

TABLE 54

SCHNEIDER INDEX: POINTS FOR GRADING CARDIO-VASCULAR CHANGES

A. Reclining Pulse Rate		B. Pulse Rate Increase on Standing				
Rate	Points	0–10 Beats Points	11–18 Beats Points	19–26 Beats Points	27–34 Beats Points	35–42 Beats Points
50–60	3	3	3	2	1	0
61–70	3	3	2	1	0	−1
71–80	2	3	2	0	−1	−2
81–90	1	2	1	−1	−2	−3
91–100	0	1	0	−2	−3	−3
101–110	−1	0	−1	−3	−3	−3

C. Standing Pulse Rate		D. Pulse Rate Increase Immediately after Exercise				
Rate	Points	0–10 Beats Points	11–20 Beats Points	21–30 Beats Points	31–40 Beats Points	41–50 Beats Points
60–70	3	3	3	2	1	0
71–80	3	3	2	1	0	0
81–90	2	3	2	1	0	−1
91–100	1	2	1	0	−1	−2
101–110	1	1	0	−1	−2	−3
111–120	0	1	−1	−2	−3	−3
121–130	0	0	−2	−3	−3	−3
131–140	−1	0	−3	−3	−3	−3

E. Return of Pulse Rate to Standing Normal after Exercise		F. Systolic Pressure, Standing, Compared with Reclining	
Seconds	Points	Change in	Points
0–60	3	Rise of 8 or more	3
61–90	2	Rise of 2–7 .	2
91–120	1	No rise . . .	1
After 120 : 2–10 beats above normal	0	Fall of 2–5 . .	0
After 120 : 11–30 beats above normal	−1	Fall of 6 or more	−1

nothing, but it illustrates forcibly the physician's inability to conceive of individual differences and the correlation of individual differences.

The above standard is true of adult men. Before the Schneider index can be used in school, it must be tried out extensively in schools, and norms established on both boys and girls of different ages.

Other measures of the cardiovascular system have been devised by Crampton and Barringer, but they will not be discussed here.

Anthropometric Measurement

Anthropometric measurement of school children has had a long and honorable history. Data on the height of children were assembled in this country on a large scale by Bowditch,[1] Peckham,[2] Porter,[3] West,[4] and Boas.[5] Other countries were assembling data on a large scale in the same period. Over 40,000 children were measured anthropometrically by 1883 in England by the Anthropometric Committee of the British Association, Sir Francis Galton, chairman. In many cases, these data, including a range of anthropometric measurement, were assembled mainly for their scientific interest and value. The facts of physical development were well known at the opening of the twentieth century. They have been admirably summarized and extended by Bird T. Baldwin.[6] There does not seem to be much excuse for an extensive compiling of anthropometric data at the present time in any one school. Usually such a practice shows the influence of

[1] Bowditch, H. P. — "The Growth of Children."

[2] Peckham, C. W. — "The Growth of Children."

[3] Porter, W. T. — "The Growth of St. Louis Children."

[4] West, G. M. — "Worcester School Children — Growth of Head, Body, Face," *Science*, 21 : 2–4, 1893.

[5] Boas, F. — "The Growth of Toronto Children," *British Association for the Advancement of Science*, 6 : 443–449, 1897.

[6] Baldwin, B. T. — *The Physical Growth of Children from Birth to Maturity.*

tradition. There are two sets of data, however, which seem to be of significance in estimating the general health of a child.

HEIGHT-WEIGHT INDEX

One of the most significant measures of general health is the height-weight index. The state of nutrition is probably as important an index of general health as any one item among growing children and can be easily determined by a single calculation from height and weight. Newman says:

> Sound nutrition is a general physiological condition which connotes a healthy body in all respects and the good tone and health of its various constituent parts, its brain and nervous system, its muscular, digestive, circulatory, and lymphatic systems.

The state of nutrition may be determined from measures of height and weight and the use of the following table prepared by Dr. Thomas D. Wood and B. T. Baldwin.

In using the table first enter the column that corresponds to the age of the child being measured. Follow down the column to the line which is the same as the child's height in inches according to the figures in the left-hand column. The figure found indicates the normal weight for the age and height given. Take the difference between the child's weight and the weight given in the table for a measure of nutrition. If the child's weight is greater, call the difference plus; if less, minus. Malnutrition is arbitrarily defined as a minus difference of ten pounds. W. R. P. Emerson [1] defines malnutrition as 7 per cent under weight. Baldwin [2] gives rules and cautions for standardizing the taking of height and weight.

> The abuses of a table center in main around the instruments used in measuring, the methods of measuring and recording, and the recording

[1] EMERSON, W. R. P. — *Nutrition and Growth in Children*, p. 14, 1922.

[2] BALDWIN, B. T. — "The Use and Abuse of Weight-Height Tables as Indices of Health and Nutrition," *Journal of the American Medical Association*, 82: 1–4, Jan. 5, 1924.

TABLE 55

WEIGHT-HEIGHT-AGE TABLE FOR BOYS OF SCHOOL AGE[1]

Height (In.)	Average Weight for Height (Lbs.)	5 Years	6 Years	7 Years	8 Years	9 Years	10 Years	11 Years	12 Years	13 Years	14 Years	15 Years	16 Years	17 Years	18 Years	19 Years	Height (In.)
38	34	34	34														38
39	35	35	35														39
40	36	36	36														40
41	38	38	38	38													41
42	39	39	39	39	39												42
43	41	41	41	41	41												43
44	44	44	44	44	44												44
45	46	46	46	46	46	46											45
46	48	47	48	48	48	48											46
47	50	49	50	50	50	50	50										47
48	53		52	53	53	53	53										48
49	55		55	55	55	55	55	55									49
50	58		57	58	58	58	58	58	58								50
51	61			61	61	61	61	61	61								51
52	64			63	64	64	64	64	64	64							52
53	68			66	67	67	67	67	68	68							53
54	71				70	70	70	70	71	71	72						54
55	74				72	72	73	73	74	74	74						55
56	78				75	76	77	77	77	78	78	80					56
57	82					79	80	81	81	82	83	83					57
58	85					83	84	84	85	85	86	87					58
59	89						87	88	89	89	90	90	90				59
60	94						91	92	92	93	94	95	96				60
61	99							95	96	97	99	100	103	106			61
62	104							100	101	102	103	104	107	111	116		62
63	111							105	106	107	108	110	113	118	123	127	63
64	117								109	111	113	115	117	121	126	130	64
65	123								114	117	118	120	122	127	131	134	65
66	129									119	122	125	128	132	136	139	66
67	133									124	128	130	134	136	139	142	67
68	139										134	134	137	141	143	147	68
69	144										137	139	143	146	149	152	69
70	147										143	144	145	148	151	155	70
71	152										148	150	151	152	154	159	71
72	157											153	155	156	158	163	72
73	163											157	160	162	164	167	73
74	169											160	164	168	170	171	74

Age — Years		6	7	8	9	10	11	12	13	14	15	16	17	18	19
Average Height (inches)	Short	43	45	47	49	51	53	54	56	58	60	62	64	65	65
	Medium	46	48	50	52	54	56	58	60	63	65	67	68	69	69
	Tall	49	51	53	55	57	59	61	64	67	70	72	72	73	73
Average Annual Gain (lbs.)	Short	3	4	5	5	5	4	8	9	11	14	13	7	3	
	Medium	4	5	6	6	6	7	9	11	15	11	8	4	3	
	Tall	5	7	7	7	7	8	12	16	11	9	7	3	4	

[1] Printed by the American Child Health Association and reprinted by their permission.

TABLE 56

Weight-Height-Age Table for Girls of School Age

Height (In.)	Average Weight for Height (Lbs.)	5 Years	6 Years	7 Years	8 Years	9 Years	10 Years	11 Years	12 Years	13 Years	14 Years	15 Years	16 Years	17 Years	18 Years	Height (In.)
38	33	33	33													38
39	34	34	34													39
40	36	36	36	36												40
41	37	37	37	37												41
42	39	39	39	39												42
43	41	41	41	41	41											43
44	42	42	42	42	42											44
45	45	45	45	45	45	45										45
46	47	47	47	47	48	48										46
47	50	49	50	50	50	50	50									47
48	52		52	52	52	52	53	53								48
49	55		54	54	55	55	56	56								49
50	58		56	56	57	58	59	61	62							50
51	61			59	60	61	61	63	65							51
52	64			63	64	64	64	65	67							52
53	68			66	67	67	68	68	69	71						53
54	71				69	70	70	71	71	73						54
55	75				72	74	74	74	75	77	78					55
56	79					76	78	78	79	81	83					56
57	84					80	82	82	82	84	88	92				57
58	89						84	86	86	88	93	96	101			58
59	95						87	90	90	92	96	100	103	104		59
60	101						91	95	95	97	101	105	108	109	111	60
61	108							99	100	101	105	108	112	113	116	61
62	114							104	105	106	109	113	115	117	118	62
63	118								110	110	112	116	117	119	120	63
64	121								114	115	117	119	120	122	123	64
65	125								118	120	121	122	123	125	126	65
66	129									124	124	125	128	129	130	66
67	133									128	130	131	133	133	135	67
68	138									131	133	135	136	138	138	68
69	142										135	137	138	140	142	69
70	144										136	138	140	142	144	70
71	145										138	140	142	144	145	71

Age — Year		6	7	8	9	10	11	12	13	14	15	16	17	18
Average Height (inches)	Short	43	45	47	49	50	52	54	57	59	60	61	61	61
	Medium	45	47	50	52	54	56	58	60	62	63	64	64	64
	Tall	47	50	53	55	57	59	62	64	66	66	67	67	67
Average Annual Gain (lbs.)	Short	4	4	4	5	6	6	10	13	10	7	2	1	
	Medium	5	5	6	7	8	10	13	10	6	4	3	1	
	Tall	6	8	8	9	11	13	9	8	4	4	1	1	

and interpretation of data. When weight is to be taken, the scales should be standardized and balanced each day. They should be of recognized scientific make, graduated into one eighth or one sixteenth pounds if the English system is used. I, myself, have not found an accurate spring balance scale. Shoes and sweaters or coats should always be removed before weighing. The weight of clothing varies somewhat with sex, height, weight and season, and to some extent with fashion and social status.

Most of the standing height measurements taken by nurses and nutrition workers are inaccurate, and vitiate the use of standards. Measurements taken by means of narrow rods, especially on weighing scales, are nearly always inaccurate. This measurement should be taken with an accurately graduated measuring scale against a flat surface at least 12 inches wide, with a flat square at least 4 inches wide and 6 inches deep. A uniform position is very important. The child stands straight with shoes off, heels together, and heels, buttock, upper part of back and head against the board to which the scale is attached. The arms are held at the side in a natural position, and the head is in such a position that the visual axis is horizontal. The square may be held in either hand of the examiner, and the reading taken from the opposite margin of the scale. The square is brought down firmly on the top of the child's head, with sufficient force to make him feel the contact.

CHEST GIRTH AND CHEST CAPACITY

Two other anthropometric measurements which are probably worth making are the chest girth and chest capacity. Seaver[1] gives the following directions for taking chest girth and capacity:

Girth of lower chest (9th rib), contracted and expanded. — These measurements are taken over the ninth rib where it crosses the mammary line. Pass the tape around the chest midway between the lower border of the large pectoral muscle and the lowest free rib. Take the first measurement after a full inspiration and the second after a full expiration.

Capacity of lungs. — This test is made with a wet spirometer with a capacity of 400 cubic inches (6.56 liters). Hutchinson's wet spirometer is considered the most reliable instrument of the kind. Directions: Fill the lungs to their utmost capacity and blow slowly into the spirometer, finishing the effort with a little spurt. Note the highest figures which appear on the rising tube.

[1] SEAVER, J. W. — *Anthropometry and Physical Examination*, pp. 50–53, 1896.

Rogers[1] finds the advice "Keep on drawing in your breath longer" and "Raise your chin more" helpful. Whipple[2] gives full treatment to the method of determining capacity of lungs.

Norms are given for chest capacity by Whipple from Smedley in terms of cubic centimeters.

TABLE 57

CHEST CAPACITY, NORMS IN CUBIC CENTIMETERS (FROM SMEDLEY)

AGE	BOYS	GIRLS	AGE	BOYS	GIRLS
6	1023	950	13	2108	1827
7	1168	1061	14	2395	2014
8	1316	1165	15	2697	2168
9	1469	1286	16	3120	2266
10	1603	1409	17	3483	2319
11	1732	1526	18	3655	2343
12	1883	1664			

The ratio of chest capacity (in cu. cms.) to weight (in kgs.) is called the vital index. Norms for the vital index are:

TABLE 58

VITAL INDEX: NORMS (FROM SMEDLEY)

AGE	VITAL INDEX	AGE	VITAL INDEX
9	69.32	15	63.18
10	69.37	16	65.94
11	69.18	17	65.77
12	67.51	18	64.28
13	66.75	19	66.22
14	64.07	20	65.01

[1] Rogers, F. R. — *Physical Capacity Tests in the Administration of Physical Education.*

[2] Whipple, G. M. — *Manual of Mental and Physical Tests, Part I*, pp. 70–74, 1910.

Strength Measurement

Rogers [1] has done the best work in determining the value of various strength indices. Grip of the right or left or preferred hand is the one measure of strength that seems most indicative of athletic ability. Grip is taken with the manometer. Rogers advises using the ordinary elliptical type manometer. Trials during which the instruments slipped in the hand were not recorded. Grip is one of the most easily determined of any of the strength indices.

TABLE 59

Grip: Norms in Kilograms (from Smedley)

Age	Boys		Girls	
	Left Hand	Right Hand	Left Hand	Right Hand
6	9.21	8.48	8.36	7.74
7	10.74	10.11	9.88	9.24
8	12.41	11.67	11.16	10.48
9	14.34	13.47	12.77	11.97
10	16.52	15.59	14.65	13.72
11	18.85	17.72	16.54	15.52
12	21.24	19.71	18.92	17.78
13	24.44	22.51	21.84	20.39
14	28.42	26.22	24.79	22.92
15	33.39	30.88	27.00	24.92
16	39.37	36.39	28.70	26.56
17	44.74	40.96	29.56	27.43
18	49.28	45.01	29.75	27.66

Rogers [2] also found strength of back and legs by means of the ordinary back and leg dynamometers. Athletic ability was predicted only slightly less by strength of back and legs than by grip.

[1] *Op. cit.*, p. 14. [2] *Op. cit.*, pp. 32–33.

Whipple[1] gives a full treatment of the strength tests in his Manual, which should be referred to. Whipple makes notes as follows concerning the grip test.

The chief objections which have been made to the employment of the hand dynamometers are (1) that it is painful, particularly if a series of grips is taken, (2) that some subjects suffer from sweating of the hands, particularly when excited, and that this causes the instrument to slip in their grasp, (3) that a wrong manner of holding the instrument may reduce the record — *e.g.*, by as much as 10 kg., (4) that, owing to the large number of muscles concerned, a lack of proper coördination in their contraction may lower the record.

The painfulness of the dynamometer can be largely eliminated by proper construction; the Smedley instrument is much better than the Collin elliptical form so commonly used heretofore. Moreover, if an extended investigation is to be undertaken, inurement to the pressure is rapidly developed (Bolton and Miller).

For the ascertainment of strength of grip, excessive perspiration can be avoided by simply drying the hands with a towel whenever necessary.

The proper holding of the instrument is also largely dependent on proper construction, and in this respect, again, the Smedley instrument, with its adjustable grip, is a distinct improvement over other forms.

The last objection is not to be seriously considered, first, because hand grip is one of the most common forms of coördinated movement and is well organized in early childhood, and second, because experience shows that most subjects can make their maximal record in three attempts at least.

Rogers[2] also used as measures of strength number of push-ups and number of pull-ups. Push-ups were performed on parallel bars and pull-ups on the rings. These tests are simple and require a minimum of equipment and show good reliability and significance as related to athletic ability. In many ways these are the most attractive of the strength tests as the records are readily obtained. Pull-ups seem to be less significant than push-ups.

Rogers[3] found that these different measures of strength are by no means identical. Grip, for instance, correlates with combined

[1] *Op. cit.*, p. 78. [2] *Op. cit.*, pp. 4, 6, 14. [3] *Op. cit.*, p. 15.

pull-up and push-up .66, while leg strength correlates with combined pull-up and push-up .59. Probably it is worth while to use all the different measures of strength described combined as a general measure of strength. For instance, Rogers found that his strength measures correlated with athletic ability as follows:

	Athletic Index
lung capacity	.59
right grip	.68
left grip	.68
back lift	.66
leg lift	.64
push-up	.67
pull-up	.59

while the multiple correlation coefficient of their combination with athletic ability is .81.

Measurement of Athletic Proficiency

Numerous scales have been devised as measures of athletic proficiency. No detailed study of relationships has been made, so that the merits of the various athletic events as measures of all-round athletic proficiency are not known. Perhaps the best procedure is to use Stecher's age norms of athletic performance, and await more exhaustive investigation of the significance of athletic stunts. These age norms increase regularly up to age 20. Age norms extending beyond 14 or 15 are always to be suspected of being influenced by selective factors. Probably it would be wiser in this case to state that these are age norms of boys and girls *in school.* One will note that these norms must be supplemented in several ways before they are of practical value. There should be norms for the various percentiles of each age group expressed not only in terms of age but in terms of height, weight, or whatever base is found to be the most highly correlated with differ-

ences of performance for any one age. This in itself is a problem for research.

The last test I shall describe is Sargent's "Physical Test of a Man." As this test involves both anthropometric measurement and athletic proficiency it does not come in any of the above four categories. This test was born out of Dr. Sargent's long and varied experience with physical education and physical measurement. One must turn to his own description of the reasoning leading up to the development of the test to see the principles back of it. Dr. Sargent[1] describes the test as follows:

> The individual to be tested stands under a cardboard disk, or paper box cover, heavy and stiff enough to hold its form, about twelve inches in diameter, held or suspended from ten to twenty or more inches above his head. He is then requested to bend forward, flexing the trunk, knees, and ankles, and then by a powerful jump upward, straightening the legs and spine, to try to touch the cardboard disk with the top of the head. Swinging the bent arms forward and upward at the time the legs, back, and neck are extended will be found to add to the height of the jump. When the disk has been placed at the highest point above the head that can be just touched in jumping, this height is measured. The difference between this height and that of the total stature is of course the height actually jumped.

The index is given by the formula

$$\text{Index} = \frac{\text{Weight} + \text{Jump}}{\text{Height}}$$

where weight is measured in pounds
jump is measured in inches
height is measured in inches

More recently L. W. Sargent[2] has found that in the Sargent

[1] SARGENT, D. A. — "Physical Test of a Man," *School and Society*, Jan. 21, 1921.

[2] SARGENT, L. W. — "Some Observations on the Sargent Test of Neuro-muscular Efficiency," *American Physical Education Review*, 29: 47–59, 1924.

test the jump is independent of both height and weight above the high school age.

Dr. Sargent provided no norms for his index, but expressed the hope that the index would prove so useful that other workers would obtain them. Unfortunately this wish will probably never be realized. Sargent's test is susceptible of great variation and probably no experimenter will try out the idea without in some way modifying it. Schwegler and Engelhardt,[1] for instance, tried out a succession of jumps in a given time period, using an apparatus which will yield accurate measurement. Here is a field for wide experimentation. The possibilities of combination of anthropometric, strength, and athletic proficiency measures seem almost unlimited.

I know of no more fascinating field for research in education than in physical education. The field is replete with problems and the data are readily available. Rogers' work, for instance, opens up many interesting problems. Some of these are:

1. To find among the various units mentioned in this chapter those that correlate most highly with such objectives of physical education as health, athletic proficiency, etc., to be determined by rating.

2. To find the intercorrelation of the several units.

3. To find the reliability of the various units. There is empirical evidence that a physician's diagnosis is extremely unreliable. Terman in *Genetic Studies of Genius*, vol. i, describes the giving of medical examinations to the gifted children used in his study. Two different physicians were employed, one examining 281 children and the other 310 children. Yet in many cases there was marked disagreement in the results. One physician reported more than five times as many deviations

[1] SCHWEGLER, R. A., and ENGELHARDT, J. L. — "A Test of Physical Efficiency: The Correlation between Results Therefrom and Results from Tests of Mental Efficiency," *American Physical Education Review*, 29: 501–505, Nov., 1924.

from the normal for the heart as reported by the other physican. Such discrepancies in the reports are so common as to lead Terman to say,

> Comparison of the records of the two physicians shows that the personal equation enters largely into the reports upon skin conditions, eye conditions (other than vision), condition of the lips, gums and tongue, abnormalities of the hard palate, deformities, heart disease, D'Espine sign, nervous symptoms, and hemoglobin. Respiration and pulse records are also subject to constant error due to the personal equation, but to a less degree. On some of the points mentioned in this paragraph one highly competent examiner will report several times as many cases of defect as another examiner of equal competence. Medical examination methods are less objective than those customarily employed by psychologists in psychometrics.[1]

For those units which show low reliability, interesting experimentation can be done on standardizing conditions of giving so as to achieve greater reliability. Both pulse and blood pressure records are considerably affected by the emotional factor. Useful experimentation should be carried on to determine means of overcoming and testing the presence or absence of the emotional influence on pulse and blood pressure.

4. Adequate norms need to be derived for the units singly and in useful combinations. These norms need to be differentiated in as many ways as there are independent variables which influence them or correlate highly with them. Norms in physical education should always be given by age. They should always be separated by sex. In many cases another variable, usually weight, but sometimes height, needs to be considered, and norms should be given in these terms. Finally, not only the central tendency but the variability should be given so that any given score may be properly placed. No measurement is readily interpreted without norms.

5. Many problems arise in connection with the technique of giving a test and the apparatus or instruments to be used. Some

[1] P. 250.

of these have been mentioned, but there are others. How reliably can the pulse be obtained by a count of 5 seconds, 15 seconds, 30 seconds? What are the relative merits from the statistical point of view of the different instruments for measuring blood pressure? What is the reliability of the hemoglobin count and what method of determination is best? What are the varieties in pulse and blood pressure at different times of day, various degrees of emotional strain before or after meals, before and after bathing, before and after smoking, before and after fatigue, etc., etc.? What is the relative reliability of the wet and dry spirometer? Which is better, the Collin or the Smedley hand dynamometer? Almost all of the athletic events have variations which can be tested experimentally. For instance, which is better in the hop, step and jump, — the unlimited or the limited run? Which gives the better results in chinning, — knuckles turned forward or backward?

6. There are a number of problems relating to practice effect, fatigue effects, warming up, effect of coaching on the strength and athletic proficiency tests.

7. Finally, there are some very neat statistical problems that are bound to grow out of much intensive work with these physical measurements. Some of these problems are: multiple correlation and regression equations with non-linear data, empirical formulæ to give highest correlations, and norms where there are several independent variables.

REFERENCES

BALDWIN, B. T. — *The Physical Growth of Children from Birth to Maturity*, UNIVERSITY OF IOWA STUDIES IN CHILD WELFARE, VOL. 1, NO. 1, 1921.

— "The Use and Abuse of Weight-Height-Age Tables as Indices of Health and Nutrition," *Journal of the American Medical Association*, 82 : 1–4, Jan. 5, 1924.

BALDWIN, B. T., and WOOD, T. D. — "Weight-Height-Age Tables," *Mother and Child*, Supplement to July, 1923, issue.

BOAS, F. — "The Growth of Toronto Children," *British Association for the Advancement of Science*, 6 : 443–449, 1897.

BOBBITT, F. — "Objectives of Physical Education," *American Physical Education Review*, pp. 229–234, May, 1921.

BOWDITCH, H. P. — "The Growth of Children," *Eighth Annual Report, Massachusetts Board of Health*, Boston; pp. 273–323, 1875.

BOWEN, W. P., and MITCHELL, E. D. — *The Practice of Organized Play*, ch. 6. A. S. Barnes & Co., New York, 1924.

COMMITTEE OF THE SOCIETY OF DIRECTORS OF PHYSICAL EDUCATION IN COLLEGES, Report of: "The Aims and Scope of Physical Education," *American Physical Education Review*, 25 : 259–261, June, 1920.

EMERSON, W. R. P. — *Nutrition and Growth in Children*, ch. 2. D. Appleton & Company, New York; 342 pp., 1922.

ENGELHARDT, J. L. — "A Test of Physical Efficiency," *Journal of Educational Psychology*, 15 : 573–578, 1924.

FOSTER, W. C. — "Test of Physical Efficiency," *American Physical Education Review*, 19 : 632–636, Nov., 1914.

GALTON, F. — "Report of the Anthropometric Committee," *British Association for the Advancement of Science*, 1881, 1883.

HEALTH EDUCATION — Report of Joint Committee on Health Problems, New York, 1924.

HETHERINGTON, C. W. — *School Program in Physical Education.* World Book Company, Yonkers, New York; 132 pp., 1922.

LEE, R. I., GEER, W. H., and BROWN, L. T. — "Bodily Mechanics in Harvard Freshmen," *American Physical Education Review*, 25 : 337–342, Nov., 1920.

PECKHAM, C. W. — "The Growth of Children," *Report of Wisconsin Board of Health*, 6 : 27–73, 1881.

PORTER, W. T. — "The Growth of St. Louis Children," *American Statistical Association Quarterly* 4, n.s. 28–34, 1894; also *Academy of Science of St. Louis*, 6 : 263–380, 1894.

ROGERS, F. R. — *Physical Capacity Tests in the Administration of Physical Education*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 173, New York; 93 pp., 1925.

SARGENT, D. A. — "The Physical Test of a Man," *School and Society*, 13 : 128–135, Jan. 29, 1921; also *American Physical Education Review*, 26 : 188–194, April, 1921.

SARGENT, L. W. — "Some Observations on the Sargent Test of Neuro-Muscular Efficiency," *American Physical Education Review*, 29 : 47–59, 1924.

SCHWEGLER, R. A., and ENGELHARDT, J. L. — "A Test of Physical Efficiency: The Correlation between Results Therefrom and Results from Tests of Mental Efficiency," *American Physical Education Review*, 29 : 501–505, Nov., 1924.

SCHNEIDER, E. C. — "A Cardiovascular Rating as a Measure of Physical Fatigue and Efficiency," *Journal of American Medical Association*, 74 : 1507–1510, May 29, 1920.

SEAVER, J. W. — *Anthropometry and Physical Examination.* New Haven; 189 pp., 1896.

SMEDLEY, F. — "Report of Department Child Study and Pedagogic Investigation," *Forty-sixth annual report, Board of Education*, Chicago, 1899–1900, 1900–1901.

STECHER, W. A. — *Theory and Practice of Educational Gymnastics.* John Joseph McVey, Philadelphia; 104 pp., 1917.

TERMAN, L. S., and Others. — *Genetic Studies of Genius, Vol. I*, ch. 9. Stanford University Press; 648 pp., 1925.

WAYMAN, A. R. — *Education through Physical Education*, chs. 5, 18. Lea & Febiger, Philadelphia; 356 pp., 1925.

WEST, G. M. — "Worcester School Children, Growth of Head, Body, Face," *Science*, 21 : 2–4, 1893.

WHIPPLE, A. M. — *Manual of Mental and Physical Tests*, Part I, Simpler Processes. Warwick & York, Baltimore, 365 pp., 1914.

WILLIAMS, J. F. — *The Organization and Administration of Physical Education*, ch. 12. The Macmillan Company, New York; 325 pp., 1923.

WOOD, T. D., and BROWNELL, C. L. — *Source Book in Health and Physical Education*, chs. 3, 18. The Macmillan Company, New York; 590 pp., 1925.

WOOD, T. D., and CASSIDY, R. F. — *The New Physical Education.* The Macmillan Company, New York; 457 pp., 1927.

CHAPTER XI

MISCELLANEOUS TESTS

The foregoing chapters have described and criticized certain outstanding tests in English, foreign language, mathematics, science, history, and physical education. Although these subjects form the backbone of the secondary curriculum, they are by no means exhaustive. It is the purpose of this chapter to discuss measurement in certain other secondary subjects and to describe the more significant tests and scales that have been constructed. The subjects to be covered are music, drawing, home economics, shorthand, and typewriting.

Music

I am not able to find a satisfactory or adequate statement of the objectives of music education in the high school. The following abbreviated list of objectives I have derived from the report of the Music Supervisor's National Conference, *Bulletin No. 1*, 1921. The standard course in music described in this report is for the elementary grades, but it is likely that objectives in the high school will merely extend these elementary school objectives.

OBJECTIVES IN MUSIC EDUCATION FOR THE HIGH SCHOOL

1. Ability in sight singing
2. Ability to sing in parts in chorus
3. Ability to sing well a certain number of songs
4. Technique on some instrument
5. Ability to play an instrument in orchestra
6. Knowledge of elements of musical notation

7. Knowledge of composers, compositions, etc.
8. Recognition of and acquaintance with certain standard compositions
9. Pleasure in good music
10. Certain abilities called for in the course in elementary harmony

Not all of these objectives are for all pupils, and some, such as those aimed at in the course in harmony, are for advanced pupils only.

TESTS OF MUSICAL CAPACITY

Tests of musical expression or appreciation have been developed to measure the native equipment of an individual. These native functions are such as do not change with experience, but simply perfect themselves with the natural maturation of the bodily functions. Kwalwasser [1] divides tests of musical capacity into three groups: 1. Tests of Sensory Discrimination. 2. Feeling Tests. 3. Motor Tests.

SEASHORE MUSIC TESTS

In the first group there are tests of the sensitivity of the learning mechanism, including such functions as (1) the sense of pitch, (2) the sense of intensity of sound, (3) the sense of time, (4) the sense of consonance, (5) the sense of rhythm. Seashore [2] and his pupils have devoted much research to the development of tests of discrimination of these different senses which play a part in music. These tests are described by Seashore in his book, *Psychology of Musical Talent.* Six of the Seashore tests have been placed on phonograph records by the Columbia Phonograph Company. These records may be played on any good standard phonograph and this provides a test that may be given to a group. The pupils should be supplied with test blanks for recording their judgments as they listen to the records.

[1] KWALWASSER, J. — "Scientific Testing in Music," *Proceedings of the Music Teachers' National Association,* pp. 155–163, 1925.

[2] SEASHORE, C. E. — *The Psychology of Musical Talent,* 1919.

Test 1 is a test of pitch discrimination. The directions are: "You will hear two tones which differ in pitch. You are to judge whether the second is higher or lower than the first. If the second is higher, record H; if lower, record L."

Gradually the two tones are brought closer together in pitch until finally an individual's answers are a matter of chance. Since a certain number of trials is allowed, the score is the number of correct responses. Tables are given for converting the per cent right into percentile ranks. Seashore states that "the sensitiveness of the ear to pitch differences cannot be improved appreciably by practice.[1]" He also states, "Knowledge of the actual, natural capacity of a child in this respect should be taken into account in determining the kind and extent of musical training the child should attempt, and other things being equal, the better the ear, the more profitable and the larger the range of possibility in musical achievement, in so far as the appreciation of pitch is concerned."[2] Seashore presents a scatter-diagram of the sense of pitch and teachers' estimates of ability in singing, but does not compute the coefficient of correlation. I have computed this coefficient from Seashore's data and find it to be .34. Even allowing for the inaccuracy of measurement, this does not seem to substantiate the foregoing statements of the prognostic value of the test.

Test 2 is a test of discrimination of intensity of sound. The directions are: "You will hear two tones which differ in loudness or strength. You are to judge whether the second is weaker or stronger than the first. If the second is stronger, record S; if the second is weaker, record W." Gradually the difference between the intensity of the two sounds is made less. The test is scored in the same way as Test 1. This test also is not affected by practice, so that it seems that the function is innate. Seashore

[1] From *The Psychology of Musical Talent*, by Carl E. Seashore, published by Silver, Burdett & Company, p. 59. [2] *Ibid.*, p. 62.

presents no objective data as to the prognostic value of this test.

Test 3 is a test of sense of time. The directions are: "You will hear three clicks marking off two intervals of time. If the second interval (that is, the time between the second and third clicks) is longer than the first interval, record L; if it is shorter, record S." Seashore does not deny that this function may be increased by training, but believes that the gain is due more to knowledge of time forms than to gain in actual capacity. Seashore states that "correlating the rating of the children in singing, as judged by their teachers, with measured capacity in time sense, we find a fairly marked tendency toward agreement," but does not state the size of the correlation, which we surmise to be very low.

Test 4 is a test of consonance. The directions are: "You will hear two combinations of two tones each; one combination is better or worse than the other in consonance (harmony). A good combination is one in which the two tones are smooth, and blend, tending to go together into one. A bad combination is just the opposite. If the second combination is better, record B; if worse, W." Training seems to give some increase in this function, although probably not to the extent generally supposed. It correlates somewhat with intelligence. Seashore does not state this test's prognostic value.

Test 5 of the Seashore series is a test of tonal memory. The directions are: "In each trial you will hear a series of tones played twice. In the second playing, one note is changed. You are to record, by number, which one was changed. In listening count mentally; for example 1, 2, in the first playing, and then likewise in the second playing, so that you may identify the one that was changed without error." The tone sequences differ in length so that this becomes a test of tone memory span.

Test 6, a new one not included in the original series, is a test of sense rhythm. The directions are: "You will hear in rapid

succession two rhythmic patterns: the second is either the same as the first or different. Listen and record either S, for same, or D, for different."

It has been claimed that the Seashore tests may be used for the discovery of musical talent; that they show the magnitude of individual differences in any one group; that they reveal differences which may not even have been suspected; that they may aid a pupil in selecting an instrument. By means of the Seashore tests the teacher may be aided in adapting the training in music to the native equipment of the child; in placing the pupil properly in his work in music in the school; in checking up on what the child is capable of doing. These tests have been advertised widely as tests of musical capacity, as undoubtedly they are, if "musical capacity" is correctly interpreted. Without the capacity which these tests measure one cannot expect to become proficient in the performance phases of musical achievement. However, the evidence that the Seashore tests are good prognostic tests of achievement in music in school does not rest on the same statistical basis as that which we demand for prediction in other school subjects. Tests of achievement in music are needed so that they may be correlated against the Seashore tests on a group of individuals who have been given extensive musical training. Only when that is done can we be certain of the value of the Seashore tests in guidance. Probably the Seashore tests need to be supplemented by intelligence tests of some sort before the best prediction on the basis of ability can be made.

There seems to be a large number of combinations of sound stimuli and verbal, vocal, or motor responses that can be used as tests. These various combinations need to be studied in order to determine their correlations and the prognostic value. At present we must take largely on faith Seashore's assertions based on his wide experience as to the prognostic value of his discrimination tests.

TESTS OF MUSICAL ACHIEVEMENT

Kwalwasser[1] divides tests of achievement into three groups: (1) knowledge or learning tests; (2) music appreciation tests; (3) performance tests. To test proficiency in singing or instrumental performance is next to impossible. Patterning after the "Gray Oral Reading Test," Hillbrand has made a "Sight Singing Test." This can be given only as an individual test. There is an elaborate system for recording errors. Unless one uses some kind of checking device the measurement of performance is reduced to unaided judgment, which is naturally subjective. Because of the nature of sound one cannot prepare a scale which can be inspected readily as one inspects a composition or handwriting scale. Perhaps the voice could be caught on the dictaphone and compared with a scaled series of records of a song, but this has not been attempted. Not much has been done in the testing of appreciation of music. Trabue[2] has investigated taste in orchestral music and has developed scales for this purpose, but his work has not had any practical outcome. We are reduced, then, to paper-and-pencil tests, which are easier to construct and which measure one's ability with the symbols of music rather than with the music itself. This is far from satisfying the objectives of music education, but at least represents the extent to which the movement of testing achievement in music has gone to-day.

KWALWASSER-RUCH TEST OF MUSICAL ACHIEVEMENT

The most representative and best of the paper-and-pencil tests is the "Kwalwasser-Ruch Test of Musical Accomplishment."[3]

[1] *Op. cit.*, p. 157.

[2] Trabue, M. R. — "Scales for Measuring Judgment of Orchestral Music," *Journal of Educational Psychology*, 14: 545–561, Dec., 1923.

[3] Published by the Extension Division of the University of Iowa (1924).

The authors are Jacob Kwalwasser and G. M. Ruch. Only one form is available. The test is useful from the fourth grade up.

The following table gives the list of sub-tests and other pertinent facts.

TABLE 60

KWALWASSER-RUCH TEST OF MUSICAL ACCOMPLISHMENT: TESTS, MAXIMUM SCORE, AND TIME ALLOWANCE

Test	Max. Score	Time Limit in Minutes
1. Knowledge of musical symbols and terms	25	5
2. Recognition of syllable names	25	4
3. Detection of pitch errors in a familiar melody	25	3
4. Detection of time errors in a familiar melody	15	3
5. Recognition of pitch names	25	4
6. Knowledge of time signatures	20	3
7. Knowledge of key signatures	30	4
8. Knowledge of note values	15	3
9. Knowledge of rest values	15	3
10. Recognition of familiar melodies from notation	50	8
	245	40

The test is either in the single-answer or multiple-choice form throughout, making it entirely objective. One may judge for himself how closely it fits the objectives of music education. I believe that most persons would agree that the test too much overweights knowledge of the symbolism of musical notation to be a really representative test of what high school musical education is striving for.

Norms are given in Table 61.

Kwalwasser has more recently compiled a test of musical knowledge, including history and biography of composers, instrumentation, and knowledge of music structure and form.

TABLE 61

KWALWASSER-RUCH TEST OF MUSICAL ACCOMPLISHMENT: GRADE NORMS AND PERCENTILE NORMS FOR THE HIGH SCHOOL

GRADE	MEAN SCORE	PERCENTILE	PERCENTILE SCORE
4	72.3	90th	223
5	86.1	80	208
6	105.1	70	197
7	120.7	60	184
8	133.7	50	171
9	150.5	40	159
10	163.1	30	142
11	171.4	20	125
12	175.1	10	104

Other tests in music have been constructed by Beach, Mosher, Gildersleeve, Torgeson, and Hutchinson, but they cannot be described here.

DRAWING

The pioneer attempt to measure ability in freehand drawing was the work of Thorndike[1] in 1913. He collected many samples of children's drawings and arranged them in order of merit. The judgments of artists, teachers of art, and students of education and psychology were used for determining the order and the difference in merit between samples used in the scale. Fifteen children's drawings chosen from Kerschensteiner's *Die Entwickelung der Zeichnerischen Begabung* were used as samples. Care was taken to select a drawing which represented as nearly as possible zero merit as the first sample in the scale.

[1] THORNDIKE, E. L. — "The Measurement of Achievement in Drawing," *Teachers College Record*, 14: 345–382, Nov., 1913.

KLINE-CAREY DRAWING SCALE

A more recent scale[1] has been constructed by L. W. Kline and G. L. Carey. They have extended their samples up to the higher ranges of quality, making the scale useful in high school.

The scales may be described by giving the authors' instructions to the pupils in collecting the original samples.

1. Draw a house.
2. Draw a rabbit.
3. Draw a figure in action (boy running).
4. Make a brush drawing of a tree.

There are four scales for each of the four subjects constructed according to the traditional methods of constructing quality scales. The unit is the P. E. of the distribution of the ratings for the various samples. The authors suggest that a pupil make a drawing of each of the four subjects and that his final rating be an average of these four ratings. There are twenty samples in the scale of drawing a house; 19 in the tree scale; 18 in the rabbit scale; and 16 in the boy running scale. Each sample has accompanying it a legend telling the pupil in what way it is better than the preceding sample. The scales as a whole have been constructed for the use of the pupils themselves. No norms of performance are given in the monograph describing the scales.

Thorndike[2] criticizes the "Kline-Carey Scale" on several counts. In the first place, the results have been left in four separate scales. In the second place, in only one of the scales (the house) is the zero point actually determined by ratings. In the other three scales the sample rated lowest by the judges was arbitrarily taken as zero, and this zero point is not con-

[1] Published by the Johns Hopkins Press, Baltimore.

[2] THORNDIKE, E. L. — "A Scale for General Merit of Children's Drawings," *Teachers College Bulletin*, Dec., 1923.

nected with the rest of the scale by ratings. In the third place, there is no attempt to secure the equality of the zero points. These criticisms are leveled against the scales as scientific instruments of measurement — they do not invalidate the use of the scales in actual school practice.

The "Kline-Carey Scale" has been subsequently (1923) revised and the features to which Thorndike objected have in part been eliminated. Each scale has a zero item obtained by ratings, but since the directions say, "Mark 0 (zero) on the back of the poorest specimen," it may be seen that the raters were not permitted to express an independent judgment as to the real value of the poorest specimen. Likewise an attempt is made to calibrate the scales by transmitting them to a single scale running from 0 to 90–95. The authors [1] state,

> The steps between samples of the other themes are more or less irregular, but the same values on each of these three scales mean practically the same distance from zero, or in other words, a drawing of a rabbit or of a dog or of a squirrel rated 70–75 on the Rabbit Scale equals in achievement a boy running or skating or what not rated around 75 on the Figure in Action Scale. Two such drawings assigned the same scale number as found on the two scale sheets in question are assumed to represent the same achievement.

When it is noted that there is no common element in any two of the four scales, it will be seen that this identity of scale values is more apparent than real. It is unfortunate, too, that a scale was chosen which bears such a resemblance to the percentage scale, inasmuch as workers in measurement are striving to free themselves from use of the percentage scale.

Art teachers naturally are averse to the mechanical features of measurement and to believing that drawing ability can be measured. But once the Thorndike Scale was made available the

[1] Kline, L. W., and Carey, G. L. — "A Measuring Scale for Free-Hand Drawing."

possibility of rating a drawing on a scale of quality was demonstrated. Kline and Carey say:

> It is doubtless true that we have acquiesced all too long in the notion that the products of drawing were due to more or less mysterious gifts independent of methodical instruction and painstaking effort. When it is once realized that drawing talent, so called, shares in equality of distribution with the gifts involved in the mastery of the other common branches, and that higher products of art are made possible by the mastery of its fundamental technique and by reducing to habit and to graphic language the fundamental art norms, the desirability of having reliable standards and norms becomes apparent.

Home Economics

OBJECTIVES

A statement of the objectives of home economics education is to be found in the committee report on the *Reorganization of Home Economics in Secondary Schools.*

> The aim of general home economics in the elementary school and high schools should be, first: To prepare the pupils for helpful and worthy membership in their present homes by establishing such standards of character as will result in consideration of the comfort and convenience of others and in willing service for the common good. To accomplish this aim it is necessary to develop skill in the use of household materials, utensils, and machinery, to inculcate such personal habits and standards as to foods, clothing, and surroundings as will insure good physical health; to train in thrift, economy, and business methods that the pupil may appreciate the problems confronting the administrator of the family income; to apply to daily life the fundamental laws of beauty of color, line, and form.

The committee has translated these aims into separate courses in which (1) the care, selection, and making of clothing; (2) selection, preparation, service, etc., of foods; (3) household economics; (4) the home and its care as regards beauty, hygiene, and sanitation are separately stressed. A balance is struck between *knowledge*, *habits*, and *skill*. Of these three, knowledge is much the

easiest to test, as it lends itself readily to paper-and-pencil testing. Skill is open to testing methods where some product or result is achieved which can be scaled to form a basis for comparison. Many skills, such as the preparation of food, are not easy to test, and no satisfactory and simple method has been devised to test such habits as system in keeping one's personal accounts. The two tests chosen to be described here are the "Home Economics Information Tests," which measure knowledge of home economics principles, and the "Murdoch Sewing Scale," which measures skill in stitching.

HOME ECONOMICS INFORMATION TESTS

The "Home Economics Information Tests"[1] were constructed in 1923 after three years' investigation by graduate students in Home Economics at Teachers College. They come in three booklets: Set I, dealing with studies which pertain to clothing; Set II, dealing with studies which have to do with foods; and Set III, covering the field of other household activities.

The items have been put into the form of statements of the multiple-choice type, the correct conclusion of any statement being indicated by underlining. The tests are intended for girls who have completed the eighth grade. A pamphlet with directions for giving and a key for scoring accompany the test.

One may grow very enthusiastic over these tests. If the girls graduating from the eighth grades of our country can be equipped with information such as is indicated by the questions in these tests, one feels that homes of the country will reach higher standards than ever before. A test of this kind sets standards for the course of study. Teachers of household arts have definite objectives to aim toward in their teaching. It is understood that the methods of teaching will have done more than simply impart information. The information alone will be of comparatively little

[1] Published by Bureau of Publications, Teachers College, Columbia University.

TABLE 62

HOME ECONOMICS INFORMATION TESTS: TESTS AND NUMBER OF ITEMS

I. *Clothing*		
Test 1	Textiles	21 items
Test 2	Construction	16 items
Test 3	Care and Repair	32 items
Test 4	Selection	35 items
II. *Foods*		
Test 1	Sources of our Common Foods	20 items
Test 2	Food Selection	15 items
Test 3	Food Preservation and Storage	20 items
Test 4	Laboratory Practices	30 items
Test 5	Food Values and Health in Meal Selection	25 items
Test 6	Food Preparation	20 items
III. *Household Activities*		
Test 1	The Girl's Bedroom	12 items
Test 2	Dining Room	21 items
Test 3	Dishwashing	11 items
Test 4	Care of the Kitchen	14 items
Test 5	Labor-saving Devices	10 items
Test 6	Home Enjoyment	9 items
Test 7	Care of Children	12 items
Test 8	Budget	8 items

value without the accompanying practical experience. One might as well try to learn to swim by reading a book on swimming. However, that does not mean that we should minimize the importance of information, always with the proviso that it should be supplemented with practical experience. What an outlook for the future homes in the United States if the domestic arts in our junior high schools could be governed by standards such as are represented by these tests!

Since many of the answers to the questions represent standards of taste or action based on the best present-day technical informa-

tion, certain of the elements may be criticized as representing neither contemporary standards held in other home-economics centers, nor standards which will be held in the future when science reverses or corrects its present findings. Such a criticism is not peculiar to this particular test. Any test in which standards of conduct are tested must necessarily represent the standards of a particular group. It is sufficient to say that this test represents the honest standards of a school which has acknowledged leadership in the field of home economics.

MURDOCH SEWING SCALE

The "Murdoch Sewing Scale"[1] was constructed in 1919 by Katharine Murdoch. The scale is useful at any point in the school system where sewing is taught. It consists of a series of fifteen photographs of sewing samples. Each sampler is shown in three views. One view gives the full-sized sampler, showing a hem, a seam, hemming, basting, running, backstitch, overcasting, and combination stitch. A second view shows the reverse side of the seam, the overcasting, basting, and backstitch. A third view shows the reverse side of the hem, the hemming stitch, basting, running, and combination stitches. Murdoch states that the range of general merit in the sewing upon the samplers is from that of an imbecile boy with very poor motor control and eye defect to that of the most expert sewer the author could find among several colleges of practical arts. All fifteen of the samplers have defined values of approximately equal steps. The samplers have been scaled according to the commonly used method which assumes that differences which are equally often noticed are equal.

Although almost any form of sewing might be measured by means of the scale, the best results are probably obtained by following the author's directions for obtaining the original samplers.

[1] Published by Bureau of Publications, Teachers College, Columbia University.

A single sampler rated by means of the scale gives an imperfect measure of any child's sewing ability. Murdoch has estimated that "if it is desired that our knowledge of her [the pupil's] sewing ability have a correlation of .9 with her real sewing ability (that ability which is the average of *all* her sewing), it is necessary that from four to five such sewing samplers be made by her on different occasions, and the average of all of them be used as the measure of her ability."

Supplementary data on the "Murdoch Sewing Scale" have been supplied by Clara M. Brown. The average correlations of scores on the scale with class marks in sewing was .533. Norms are given in the following tables:

TABLE 63

MURDOCH SEWING SCALE: NORMS

GRADE	SCORE	AGE	SCORE
6	4.6	10	2.9
7	6.0	11	5.5
8	7.6	13	6.5
9	9.0	14	7.5
10	10.8	15	8.3
1st-year college	11.5	16	9.1
2nd-year college	12.7	17	9.7
		18	10.8

More recently (1922) Murdoch has published analytic scales of sewing ability. There are five separate scales for each of the five following stitches: hemming, overcasting, running, combination, and backstitch. Each scale consists of five samples, corresponding to samples in each of the scales having the same merit. No attempt is made to cover the whole range of sewing ability. Since these analytic scales were designed for the pupils' own use,

samplers possessing little merit were not included. These analytic scales have been calibrated with the scale of general merit in sewing ability so that the units on each scale represent the same level of ability as is represented by the units on the general scale.

Stenography

OBJECTIVES

A whole phase of secondary education — the commercial curriculum — must have only brief treatment here. The Committee on the Reorganization of Secondary Education, in its report on *Business Education in Secondary Schools*, has made the following analysis of Commercial Education:

> In a general way, commercial education up to the present time has attempted to meet four distinct business needs. First, and most definite of these, is the training of stenographers; and second, is the training of bookkeepers and clerks for general office work. These two functions have hitherto been regarded as the full obligation of commercial education. Third, the need that business education has recently undertaken to meet is the training for secretarial work of those who have had a broader fundamental education and who wish to take more responsible positions than to be merely stenographers. Fourth, the need that commercial education now seeks to supply is the demand for salesmen.

For testing salesmanship or specialized secretarial duties there are almost no testing techniques, except insofar as stenography is one of the secretary's tools. There are no standardized tests of bookkeeping. Commercial arithmetic may be tested by standardized arithmetic tests. The reader is referred to manuals of elementary tests for the description of arithmetic tests.

This section is devoted to a description of tests of shorthand and typewriting which are amenable to objective testing and for which good standardized tests have been constructed.

SHORTHAND

HOKE SHORTHAND TESTS

R. E. Hoke has constructed a battery of tests for measuring different phases of ability in shorthand. Ability in shorthand may be analyzed into (1) ability to translate words into shorthand symbols, *i.e.*, the acquisition of the shorthand vocabulary, (2) ability to form the characters well, (3) ability to write in shorthand with a certain speed, (4) ability to read shorthand, *i.e.*, to translate into English. Hoke [1] constructed scales [2] for each of these abilities in the Gregg system of shorthand.

A set of ten vocabulary tests are available, each containing 150 words. This provides opportunity for testing 1,500 carefully selected words. One thousand of the words were taken directly from the Ayres list of the 1,000 most common words from the standpoint of spelling. The Ayres list was based on word counts made by Knowles, Eldridge,[3] Cook and O'Shea,[4] and additional counts made by Ayres — a total count of 368,000 words. The Gregg system of shorthand includes not only symbols for separate words but also symbols for phrases or combinations of words. A count of such phrases and combinations of words has been made on the type of material which is commonly used for dictation. Approximately 500,000 words have been read and tabulated which accounts for about 92 per cent of ordinary composition. In this count much of the material was in the form of expert notes published by the Gregg Shorthand Company. The publisher's experts have decided in cases of difference of opinion in the selection of the phrase. Six thousand four hundred eighteen different phrases were tabulated. The 500 most frequent phrases, which include 78 per cent of the total number counted, constitute

[1] HOKE, R. E. — *The Improvement of Speed and Accuracy in Typewriting.*

[2] Published by the Gregg Publishing Company (1921).

[3] ELDRIDGE, R. C. — *Six Thousand Common English Words.*

[4] COOK, W. A., and O'SHEA, M. V. — *The Child and His Spelling*, 1914.

the balance of the words used by Hoke in assembling his tests. On the basis of testing results a scale of shorthand difficulty was constructed. In each column in this scale the words are of almost equal difficulty. The difficulty in terms of per cent is given at the top of each column. These per cents fit a group that has studied shorthand a little over a year. If the group has studied only one year, 3 should be subtracted from the figures at the top of the table; if the group has studied one and a half to two years, about 3 should be added to the figures at the top of the table.

The 150 words which are included in each of the 10 forms of the Hoke test were selected at random from the 1,500 words and phrases which were chosen as described above. The difficulty of each of these 10 forms is very nearly the same. Form 1, the hardest, gives scores which are on the average 9 points lower than form 4. Any one of these tests contains easy words and harder words and may be used as a measure of achievement. The whole set of 10 tests given at the end of the year would serve as a good inventory of shorthand knowledge. No time limit is set in giving the tests. If a teacher wished to test achievement very accurately, it would be better to select the words from the scale, as in that way the level of difficulty may be controlled. That is, words may be selected from the upper end of the scale if one feels confident that the easy words at the lower end of the scale are well known. If a test is constructed from the scale, the figures along the top of the scale indicate the norm. For instance, suppose that 100 words are

No. of Words	Per Cent	No. Words Right in Each Group of 20
20	58	11.6
20	50	10.0
20	42	8.4
20	34	6.8
20	27	5.4
		42.2

selected to form the test, 20 words to be taken from each of the columns 58, 50, 42, 34, and 27.

The norm on this group of 100 words is 42.2.

For ability to write shorthand well a scale of shorthand quality of writing was constructed by Hoke. The procedure used in constructing the scale has been carefully described by him. First, a collection of 254 specimens of shorthand was made from Irving's *Sketch Book*. These specimens were copied by shorthand students, yielding 1,155 samples. Six specimens from the Gregg office, written by experts, and twenty specimens by summer-school students who had never had any instruction in shorthand, were included in the samples. One of the latter specimens was selected as a sample having zero merit. Five judges sorted these 1,155 samples into five groups. A selection was then made of 41 samples in each group, yielding 205 specimens. Twelve additional judgments were obtained on the 205 specimens and this time they were graded into 10 piles. Of these, 40 specimens were selected whose quality differed by approximately equal amounts. These 40 specimens were then engraved and sent to shorthand teachers to be ranked in order of merit. Two hundred fifty teachers responded and an average rank was obtained of the forty specimens. The difference in quality was determined by the amount of overlapping of judgments in terms of the P. E., using the usual Thorndike procedure. Sixteen of these 40 specimens subsequently were chosen in such a way that they would differ from one another in value by approximately 5 points on a scale of 100 points. This set of sixteen specimens constitutes the scale.

To test speed of writing shorthand Hoke devised a copying test using a "tri-interlinear device." In this test the copy is provided in printed longhand and also in shorthand with space included below for copying the shorthand. Hoke claims that this is the best possible way by which the function could be isolated as long eye

jumps are not required, and the copy does not depend on hearing, memory, or having to accommodate one's speed to the rate of dictation. It was found by trying this test on shorthand experts that these experts could not copy the printed text as fast as they could take from dictation. This is undoubtedly a matter of habit and is due to the fact that they learned to accommodate themselves to the habit of anticipating the spoken language. The main question concerns the degree of correlation between copying as required on the Hoke test and taking shorthand notes from dictation. On this point Hoke does not supply information. Two forms of this test, B-1 and B-2, practically equal in difficulty, are available. Two minutes are allowed in taking the tests. There seems to be a large practice effect on these tests.

To test ability to read shorthand a unique procedure was adopted. Two business letters exactly five hundred words in length were chosen. Every tenth word also has an alternate word written sometimes above it and sometimes below it — a purely chance order was chosen — to avoid any advantage from guessing. The person tested chooses the correct word. The letter itself is written in shorthand, but every tenth word with its alternate is written in English. The alternate words were chosen in order from one of the columns in the Ayres Spelling Scale in order to make the choice of alternative words purely random and particularly to avoid words where fine discrimination is needed in making the choice. The test was first tried out in English to make sure that there were no inherent difficulties in the choice of words. Three choices that seemed to offer inherent difficulties were changed before the shorthand form was constructed. The test is scored according to the key provided as "the number of words read correctly per minute (three minutes being allowed)." Hoke states the method of scoring calls for subtracting the number wrong from the number right. A correlation of .55 obtained from 41 cases was found between number

of words read aloud from a passage of shorthand notes for three minutes and the scores on these reading tests.

These tests and scales, all published by the Gregg Publishing Company, represent an excellent piece of constructive work in providing standardized tests in this specialized field. The stenography teacher would profit by using these tests in ways which are suggested in other places in this book.

GREGG DIAGNOSTIC SHORTHAND TESTS

More recently the Gregg Company has published "Diagnostic Shorthand Tests," one set of tests for each of the first four chapters of the Gregg Shorthand Manual. These tests, constructed by Ethel A. Rollinson, in general use the same forms already devised by Hoke and do not need further discussion here.

TYPEWRITING

Typewriting has long been used in studying learning, etc. Rejall and Hill, for instance, used Eliot's *Mill on the Floss* as material in a study of practice in typewriting, recording their results in minutes and seconds necessary to write a 100-word passage. Such material is far from being equal in difficulty value in different sections. Speed contests in business exhibitions have used standardized material, but material which has been selected so as to facilitate speed by eliminating long words.

BLACKSTONE TESTS OF TYPEWRITING

Standardized tests of typewriting have been constructed by E. G. Blackstone.[1] Five alternate forms are provided. The unit of measurement in this series of tests is the stroke, defining stroke as "any single movement in typing," and including the striking of any key, whether it be a letter, figure, space bar, shift key,

[1] Published by the World Book Company (1923).

tabulation key, or back spacer. It has been demonstrated experimentally that difficult strokes on the typewriter differ in difficulty, but at least they are more nearly equal than any other longer unit. The tests consist of business letters. Blackstone [1] states that these letters have been worked over until they were so worded that:

1. Each contains approximately the same number of words.
2. Each contains approximately the same number of letters.
3. Each contains approximately the same number of *e*'s, *t*'s, *a*'s, etc., with the frequency of the letters of the alphabet in conformity to the normal frequency as determined by studies of cryptography.
4. Each contains approximately the same number of letters to be struck with the right hand.
5. Each contains approximately the same number of letters to be struck with the left hand.
6. Each contains approximately the same number of carriage returns and shift-key strokes.
7. Each has the particularly long or difficult words in approximately the same positions.

" Approximately " is not defined.

It is evident that great pains have been taken to insure comparability of material in the five tests in regard to those factors which make for differences of difficulty in typing.

On page 3 of the test booklet is given the text of the passage to be copied by the pupil. Under each word is given the cumulative number of *strokes per minute* which would have been made if the test had been typed as far as and including that word. In scoring the test errors in typing are encircled. Speed and accuracy are combined in a single score by the formula:

$$\text{Score} = \frac{\text{Strokes per minute} \times 10}{\text{Errors} + 10}$$

[1] "Blackstone Stenographic Proficiency Tests," *Manual of Directions*, p. 2. Copyright 1923 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

This formula lowers the score for every additional error. The norms are given in the following table:

TABLE 64

BLACKSTONE TYPEWRITING TEST: NORMS

MONTHS OF INSTRUCTION	NORM
5	88
10	148
15	178
20	200
25	220
30	236

In the manual of directions accompanying the test percentile norms for each of the above periods may be read from a percentile graph chart. A sheet is also provided on which the score made by an individual on successive trials of the test may be plotted to indicate progress in learning.

THURSTONE EXAMINATION IN TYPING

Thurstone has constructed an Examination in Typing[1] of three parts, each of which contains novel features. First, a test of correcting a poorly typewritten copy; second, an exercise in tabulating on the typewriter; and third, a test of correcting spelling errors. In this examination the number of errors is counted in order to determine the score and the person making the fewest errors receives the highest credit.

All of the subjects discussed in this chapter offer excellent opportunity for those who wish to experiment in any phase of measurement, either in the construction of new tests or in experimental or administrative use of tests already available.

[1] Published by the World Book Company.

REFERENCES

(Music)

KWALWASSER, J. — "Scientific Tests and Measurements Applied to Music," *Music Supervisor's Journal*, May, 1924.

— "Scientific Testing in Music," *Proceedings of the Music Teachers' National Association*, pp. 155–163, 1925.

— "Music," ch. 14, of the *Third Yearbook of the Department of Superintendence of the National Education Association*, pp. 354–366, 1925.

MUSIC SUPERVISORS NATIONAL CONFERENCE — Fourteenth Annual Meeting, held in St. Joseph, Missouri, April 4–8, 1921, *Bulletin no. 1*, 1921.

SCHOEN, M. — "Recent Literature on the Psychology of the Musician," *Psychological Bulletin*, 18 : 483–489, Sept., 1921.

— "Tests of Musical Feeling and Musical Understanding," *Journal of Comparative Psychology*, 5 : 31–51, Feb., 1925.

SEASHORE, C. E. — "Avocational Guidance in Music," *Journal of Applied Psychology*, 1 : 342–348, 1917.

— *The Psychology of Musical Talent.* Silver, Burdett & Company, Boston; 288 pp., 1919.

— "Recent Progress in the Psychology of Musical Talent," *Music Teachers' National Conference Proceedings*, pp. 158–165, 1920.

— *A Survey of Musical Talent in the Public Schools*, STUDIES IN CHILD WELFARE, VOL. 1, NO. 2, University of Iowa, Iowa City; 36 pp., 1920.

TRABUE, M. R. — "Scales for Measuring Judgment of Orchestral Music," *Journal of Educational Psychology*, 14 : 545–561, Dec., 1923.

(Drawing)

KLINE, L. W., and CAREY, G. L. — *A Measuring Scale for Free-Hand Drawing, Part I, Representation*, THE JOHNS HOPKINS UNIVERSITY STUDIES IN EDUCATION, NO. 5, The Johns Hopkins Press, Baltimore, 61 pp. and copies of scales, 1922; also *A Revision of the Original Scale*, JOHNS HOPKINS UNIVERSITY STUDIES IN EDUCATION, NO. 5A.

THORNDIKE, E. L. — "The Measurement of Achievement in Drawing," *Teachers College Record*, 14 : 345–382, Nov., 1913.

— "A Scale for General Merit of Children's Drawings," TEACHERS COLLEGE BULLETIN, FIFTEENTH SERIES, NO. 6, Dec., 1923.

(Home Economics)

BROWN, C. M. — "Investigations Concerning the Murdoch Sewing Scale," *Teachers College Record*, 23 : 459–470, Nov., 1922.

Brown, C. M. — "What Educational Measurements Can Do for Home Economics," *Journal of Home Economics*, 16 : 191–196, April, 1924.

— "Construction and Use of Information Tests in Home Economics," *Journal of Home Economics*, 16 : 251–256, May, 1924.

Cooley, A. M., and Reeves, G. — "Some Investigations concerning the Use of Certain Home Economics Information Tests," *Teachers College Record*, 24 : 374–392, Sept., 1923.

Murdoch, K. — *The Measurement of Certain Elements of Hand Sewing*, Teachers College Contributions to Education, no. 103, Columbia University, New York, 1919.

— "A New Analytic Sewing Scale," *Teachers College Record*, 23 : 453–458, Nov., 1922.

(Stenography)

Hoke, R. E. — *The Improvement of Speed and Accuracy in Typewriting*, The Johns Hopkins University Studies in Education, no. 8, The Johns Hopkins Press, Baltimore, 1922.

CHAPTER XII

HOW TO HANDLE THE RESULTS OF TESTING

The foregoing chapters have described intelligence tests and standardized achievement tests. We are now ready to proceed with the statistical treatment of test results. Let us imagine that we have a set of tests which have been scored. The results are in a raw state: they are like the raw products which come into a factory. Now let us consider the methods of whipping these results into a form in which they are most easily interpreted. The mind cannot assimilate the meaning hidden in the figures as the results are first crudely put down. We need to learn how to tabulate the results of testing in an orderly way, and to know how to manipulate the results of the class as a whole so that the general condition of the class may be seen at a glance.

In Appendix A are the results of intelligence testing in a private secondary school using the "Terman Group Test of Mental Ability." Column 1 gives the ages of the pupils in years and months, Column 2 the pupils' initials, and Column 3 the scores on the tests.

These test results will serve as a text for our discussion of the handling and interpretation of test results. They are typical raw material.

Turn to the results of the intelligence tests for the first-year boys on page 546. What is the highest score? What is the lowest score? What is the highest score for the second-year boys? What is the lowest score? In getting the answers to these questions you have had to run your eye up and down the column until you have eliminated all scores except the highest

and the lowest. This operation cannot be performed very efficiently, because the scores are not conveniently arranged to facilitate this. The first task one must learn to do is to rearrange the scores in a way that will make them more easily interpreted. Again compare the intelligence scores for the first-year pupils with the intelligence scores for the second-year pupils. Which group did the better? In this case one can easily see that there are no scores below 100 and there are more scores of 190 or above in the second-year group, so that the second-year boys are presumably better, but just how much better is not clearly evident. Our second problem is to find a single figure which we may take as representative of the grade ability as a whole and which we may use to compare the ability of one grade with another or the ability of the grade with the norm.

I. Making a Distribution of Scores

One could make a distribution of scores in arithmetic for the first-year pupils as follows:

Scores
189
187
176
170
164
161
etc.

Finish this column on a separate sheet of paper for yourself. One way of telling whether you have made no error is to count your figures and see that you have 26 scores listed. This makes a long, awkward column which can only be obtained in an awkward way. A better way would be as follows:

TABLE 65

ILLUSTRATING THE CONSTRUCTION OF A FREQUENCY DISTRIBUTION

Intelligence — First Year Boys

SCORES	F	F
181–189	|||	3
172–180	|	1
163–171	||	2
154–162	||||	4
145–153		0
136–144	||	2
127–135	|||	3
118–126	||	2
109–117	|	1
100–108	|||	3
91–99	|	1
82–90	||	2
73–81	|	1
64–72	|	1
		$26 = N = \Sigma f$

This is obtained in the following manner: After reading the 64 at the bottom of the column of intelligence scores of the first-year boys, make a tally opposite the scores 64–72; then a tally opposite the scores 163–171; another tally opposite 136–144, etc. When you come to the fifth tally of any score, make it by crossing the four already entered, thus ~~||||~~. This compact bundle of little marks is easily interpreted by the eye to stand for 5. Notice that the columns are headed. One should always label the columns in making a frequency distribution. f stands for frequency and indicates the number of pupils making a given score. These tally marks may easily be transmuted into figures. The total in the f column should equal 26. This is labeled N or Σf. N stands for *number*, Σ stands for *the sum of;* and Σf is read "the sum of the frequencies."

Notice that the scores are grouped. In grouping one ought to make sure there are at least twelve classes or groups. To deter-

mine the size of the class interval, a good plan is to subtract the lowest score from the highest score and divide the difference by twelve. In this instance 64 from 189 is 125; 125 divided by 12 is 10 and a fraction. We choose as our class interval the nearest odd integer below this quotient. In this case the nearest odd integer below 10 is 9, hence we shall make our class interval a step of 9.

Notice that the class intervals are not labeled as follows:

181–190
172–181
163–172
154–163

If the class intervals were labeled in this way, one would not be able to tell in which line to place a score of 181. It might well go in either the first or second line. But with class intervals labeled as we have labeled them in Table 65, there is no chance of getting a score in the wrong place.

We now have a frequency distribution which is a convenient display of the scores, giving them in order and the number of individuals making each score. From such a frequency distribution one can readily see not only the highest and lowest scores, but what is more important, the score around which the individuals tend to cluster.

II. The Arithmetic Mean

The arithmetic mean is nothing except the old-fashioned average. It may be easily found on the original entry sheets in any class by adding the scores and dividing by N, the number of individuals taking the test. For the first-year boys this is 3,469 ÷ 26 or 133.4. It is called the arithmetic mean to distinguish it from other means, notably the geometric mean. When one has a frequency distribution, it is not possible to add the scores and divide by the number; consequently another method must be resorted to in

finding the arithmetic mean. This method is really a very short one, and for a large number of items is even more expeditious than the method which you are accustomed to. This will be referred to as the short method. The arithmetic mean of these data by the short method is 132.7. The reason that there is a difference in the value of the arithmetic mean found by the two methods is that when the scores are grouped in a frequency distribution there is some distortion of their actual values.[1]

The arithmetic mean offers a new objective in the giving of examinations. In the old-fashioned examination the teacher's objective was to get as many 100%'s as possible and to have as few failures as possible. The arithmetic mean gives us a convenient single figure to use which may be taken as representative of the whole class. One may compare the mean of a class with the mean of other classes on the same test in other schools and in other cities. The objective of a teacher should be to raise the arithmetic mean of the class to whatever goal is set. The goal thereby becomes objective and the knowledge of whether it is attained or not is objective. To be more concrete, how shall we interpret this average score of 132.7 which we have found for this first-year class? Referring to the norms for the Terman test given in Chapter IV, one sees that on the average in a ninth-year class 50% of the children make a score above 104. This class, then, with whose scores we have been working is well above this average. As a matter of fact, our class has an average intelligence exceeded by about only 22% of the first-year high school population as a whole.

The median is another common measure of the central tendency. It is not the same as the arithmetic mean, although actually it usually differs very little from the arithmetic mean. It is not necessary to know how to find both of these measures. The arithmetic mean suffices as an average in all practical work. It

[1] The computation of the arithmetic mean by the short method is described in Appendix A.

does little harm practically to compare an arithmetic mean with the median.

III. The Standard Deviation

We have told only half the story about any set of scores when we have given the average of the scores. Two classes might have the same arithmetic mean and yet be quite different in their nature. How this can be may be best illustrated by giving two distributions. The first distribution is the same as that given in Table 65. The second distribution is hypothetical.

TABLE 66

Illustrating Two Distributions Having the Same Mean but Different Variabilities

Scores	F	F
181–189	III	
172–180	I	
163–171	II	
154–162	IIII	
145–153		III
136–144	II	𝍸 II
127–135	III	𝍸 𝍸 I
118–126	II	IIII
109–117	I	I
100–108	III	
91–99	I	
82–90	II	
73–81	I	
64–72	I	
	26	26

These two distributions have approximately the same means, but one distribution has a much greater spread than the other. The scores in the second distribution are bunched more closely together. A class represented by the second distribution would probably be much easier to teach than a class represented by the

first distribution. In the second class the pupils would be more homogeneous with regard to their intelligence. A teacher of the first class would find that if she were teaching pupils who made scores of 160, 170, or 180, she would be shooting quite over the heads of pupils who made scores of 70, 80, or 90, or if she devoted her attention to the two lowest pupils, the brighter pupils would become restless and troublesome from having too little to command their attention. A teacher of the second class would have an easier task because her pupils are more nearly of the same ability.

Our problem is to find a unit or number which will serve as a measure of the spread or variability of the group just as the arithmetic mean describes the central tendency. By use of the mean and this new measure we can give a fairly complete description of a distribution.

One simple measure of deviation is known as the *average deviation.* By this is meant the average of the differences between each score and the mean of the group. For example, in our problem, the difference between the highest score (189) and the mean (133.4) is 55.6. The next difference (187–133.4) is 53.6. If we take the average of all these deviations (55.6, 53.6, etc.), we have the average deviation.

The unit which we are about to describe is the one most commonly used and is called the *standard deviation.* It is abbreviated by S. D. or σ (sigma). In finding the standard deviation one first finds the deviations from the mean and then squares these deviations ($55.6^2 = 3091.36$; $53.6^2 = 2872.96$); second, one sums these squared deviations; third, one finds the average of the sums of these squared deviations by dividing by N, and fourth, one finds the square root of the average.

The standard deviation has also been called root-mean-square deviation. This name is also a succinct description of how the standard deviation is found. First find the squares of the deviations, then find the mean of these, and finally take the square root.

The standard deviation of our problem has been found to be 34.6.

It is difficult to state the significance of the standard deviation 34.6 in simple English. The best that can be done is to say that it is a measure of the dispersion or scattering of this class in age. Another class bunched more closely around the mean, 132.7, in age would have an S. D. less than 34.6. Another class that was even more heterogeneous or scattered would have an S. D. greater than 34.6. Only further statistical study of the properties of the S. D. will give it more meaning. It does not have any simple meaning such as the arithmetic mean has. We will use the S. D. in a practical way in future chapters.[1]

Just as there are other measures of central tendency beside the arithmetic mean so there are other units commonly used in statistics to measure variability. One of these is the semi-interquartile range, whose name only I am giving here. One should be aware that there are other measures, but for the purpose of this book it is not necessary to be familiar with more than one.

IV. Types of Curves Found in Educational Distributions

This is a good place to note the general shape of the distribution that we commonly get from educational data. When one looks at the frequency distributions on page 236 of this chapter, he will see that they both have certain general characteristics. Most of these scores are concentrated somewhere in the middle of the distribution. The distribution tapers off at each end so that there are only a few very high scores and a few very low scores. This tapering off is fairly symmetrical on both sides of the mean. Distributions from educational data so uniformly give this same type of curve as to make one believe that there is something almost

[1] The computation of the standard deviation from a frequency distribution is also described in Appendix A.

necessary about it. In fact the distributions so generally take this shape that educators commonly assume that it is this shape when they know nothing else about it, and this is the safest assumption that can be made in lieu of further knowledge. These distributions bear close resemblance to a mathematical curve known as the "normal probability curve" or the "curve of error." Commonly one hears statements that would make him think that distributions of educational data are of a shape quite different from the normal distribution curve. The majority of children in any group are average children, only a few are superior and only a few are inferior. This is true not only of intellectual traits but also of moral traits. Children are not either good or bad; they are average, and only the exceptional child may be classed as very good or very bad.

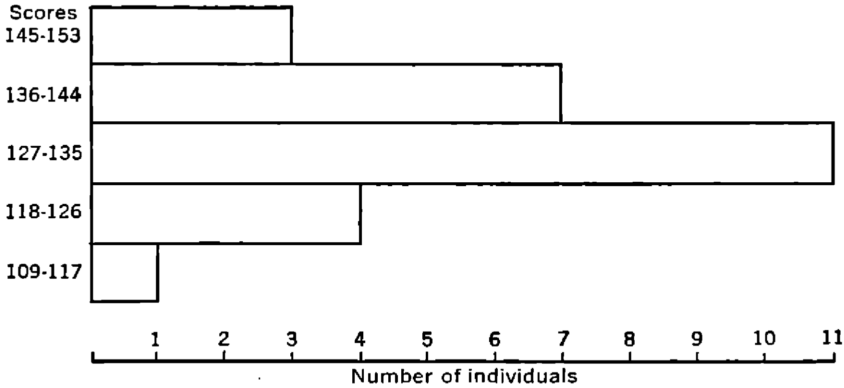


FIGURE 2.—Distribution of frequencies given in Table 66

Figure 2 is the graph of a second set of frequencies on page 236 of this chapter. As you will see, the scores run up and down as in the distributions themselves and the length of the column indicates the number of individuals obtaining a certain score. Each square quarter inch represents a certain individual.

Figure 3 is the graph of this same distribution turned over so that the scores run along the base. Again the height of a

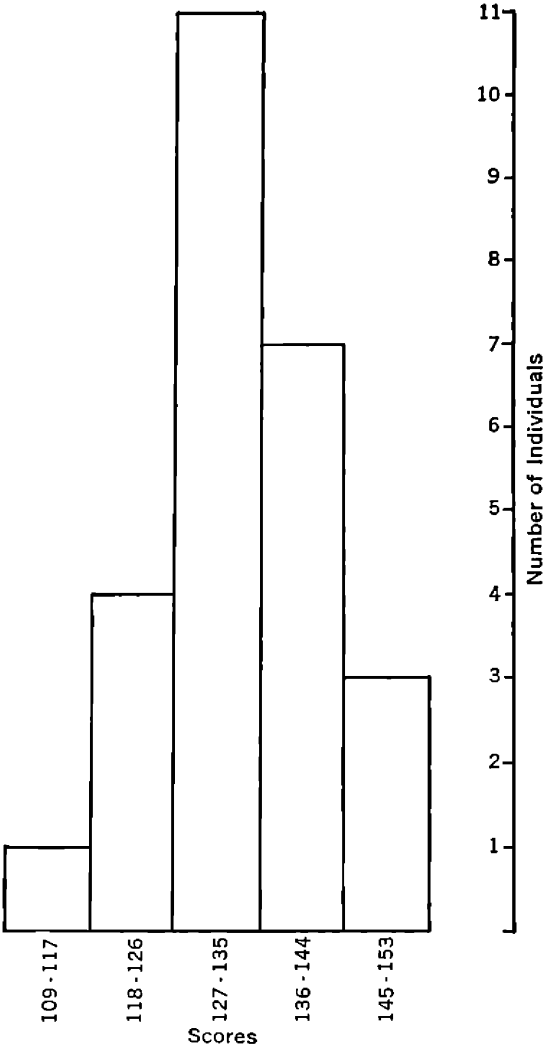


FIGURE 3.—Distribution of frequencies given in Table 66

column represents the number of individuals obtaining a certain score.

Figure 4 is a graph which is called the frequency polygon. It represents the same distribution as that represented by figures 2 and 3, except that the tops of the columns are connected by straight lines in order to give the idea of continuity.

Figure 5 is a normal probability curve drawn to fit the same data. Whenever you see it you will know that it is an ideal representation of a frequency distribution. The scores run from low to high from left to right along the base, and the height of the curve at any point indicates roughly the number of individuals who would obtain that score.

Figure 6 is the type of curve that we almost never find arising from educational data. It was drawn here to illustrate the point

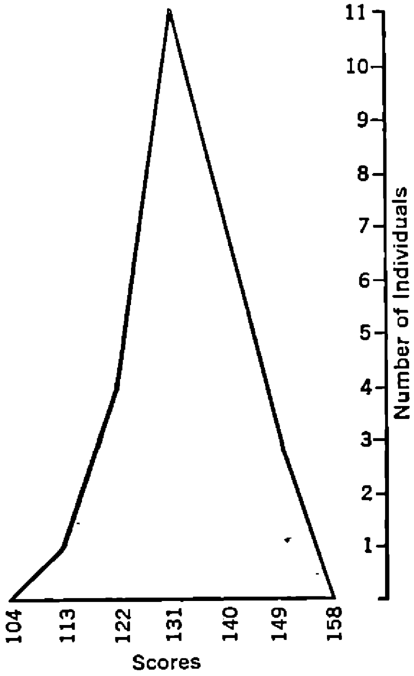


FIGURE 4. — Frequency polygon of data given in Table 66

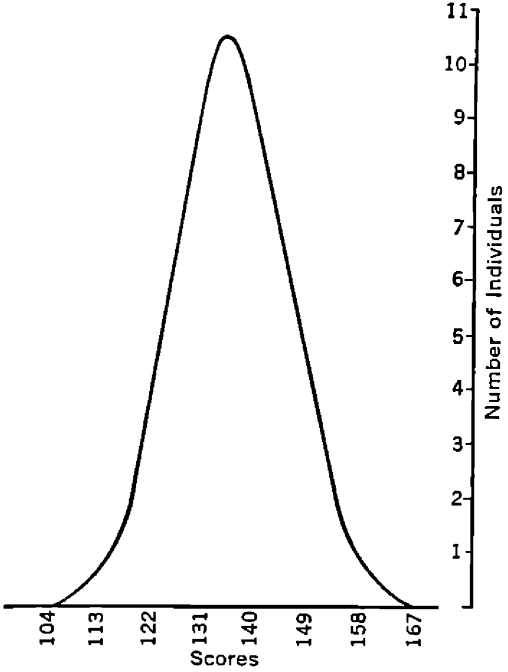


FIGURE 5. — Normal probability curve designed to fit data of Table 66

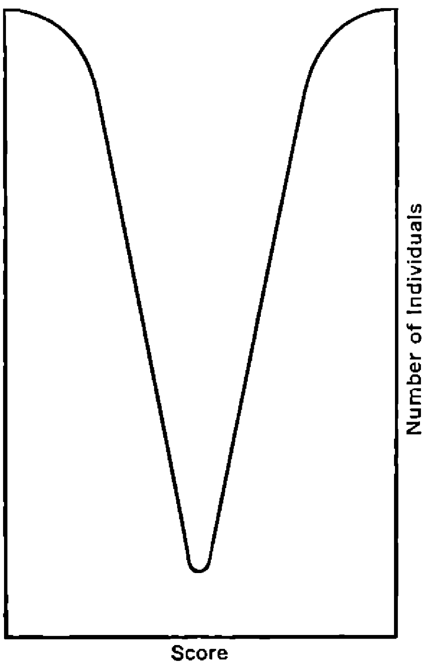


FIGURE 6. — Bi-modal frequency curve

made on the preceding page. Careless conversation would often lead one to believe that a class was distributed according to Figure 6 with many good and poor pupils and a few average pupils. As a matter of fact, as we have before said, most distributions obtained from classroom testing tend to assume the shape of Figure 5.

Some of the properties of Figure 7 may be explained in terms of the standard deviation.

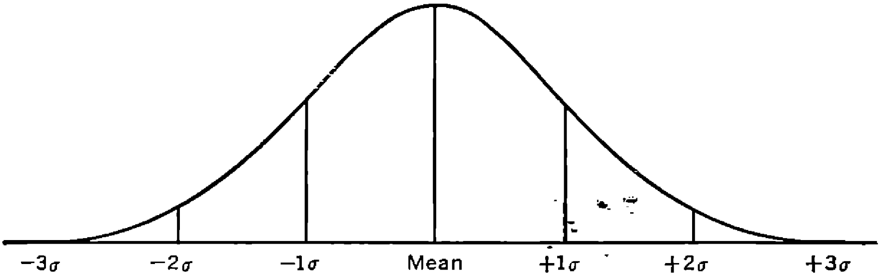


FIGURE 7. — Normal probability curve

In the normal distribution 34% of the cases lie between the mean and + 1σ. Between + 1σ and + 2σ lie 14% of the cases. Between + 2σ and + 3σ lie 2% of the cases. To the right of + 3σ lie only 0.13 of 1% of the cases or 1.3 out of 1,000.

To put this differently, between − 1σ and + 1σ lie 68% of the cases, between − 2σ and + 2σ lie 95% of the cases, whereas only 0.27 of 1% of the cases or 2.7 out of 1,000 lie outside of 3σ on either end.

3σ is therefore the practical limit of the extent of any distribution, and with a class of under 40 pupils it is only by chance that any pupil exceeds 2.5σ from the mean of the group.

Let us make this more concrete by putting it in terms of the distribution we have been working with. The mean was 132.7; the standard deviation was 34.6. 1σ above the mean is 167.3, 2σ above the mean is 201.9, and 3σ above the mean is 236.5. 1σ below the mean is 98.1, 2σ below the mean is 63.5, and 3σ below

the mean is 28.9. It is seen that in this instance all the cases are included between -2σ and $+2\sigma$. This is exceptional, and 3σ is taken as the practical limits of a distribution on each side of the mean in general. If one tries to fit the percentages given above to our problem, it will be seen that the agreement is very poor.

TABLE 67

DISTRIBUTION OF SCORES ON THE TERMAN GROUP TEST OF MENTAL ABILITY IN A PRIVATE SCHOOL

Score	Boys				Girls			
	Class				Class			
	I	II	III	IV	I	II	III	IV
200–209				|				
190–199		||	||	||		|	||	||
180–189	|||	||	𝍸	𝍸 ||||				𝍸 ||
170–179	||	|	|||	|||	||||	|||	||||	|||
160–169	||	||||	𝍸 ||	|	||	𝍸 |	𝍸 |	𝍸 |||
150–159	|||	||	||	||||	||	𝍸 |||	𝍸	𝍸 ||
140–149	|	||||	||	||	||	𝍸 ||	||	||
130–139	|||	𝍸	|	|	|||	||||	𝍸 ||	||||
120–129	|||	|	|	|	||||	||||	𝍸	|
110–119	|	|			|	𝍸	||||	|
100–109	|||	|			|||	|||	|||	|
90–99	||				𝍸	𝍸	|||	|
80–89	||				|			
70–79					|||			
60–69	|				||			
50–59					|			
40–49								
30–39					|			
Totals	26	23	23	24	34	46	41	37

The percentages are for the smooth mathematical curve and do not fit any particular case. However, they fit particular cases in general better than any other set of percentages which might be given.

Table 67 pictures the distributions of the intelligence scores for the classes of a secondary school. There are many interesting things to note in this figure. First of all, one should note the regular and gradual progress in intelligence through the classes in the school. As we should expect, pupils in each grade are on the average better than those in the preceding grade. Part of this may be due to selection and elimination of the pupils and part to actual improvement. The second thing to note is the wide variability in any class as compared with the difference in average score between one class and the next. Where the difference between the average of one class and the next is about 15 or 20 points, the range of scores in any one class is some 100 points more or less. Another way of looking at this is to note the great overlapping between classes. The difference between one class and the next is not clean cut. As a matter of fact the difference between classes III and IV appears very slight. Three or four pupils in class I are better than the average pupil in class IV and three or four pupils in class IV are not so good as the average pupil in class I. A picture of this kind shows us that there is no clean-cut distinction in ability between pupils of one class and another in the high school. There is great overlapping of ability. Were it not for the matter of prerequisites in many subjects, pupils in the junior and senior classes might be interchanged with no apparent difference in their ability to do the work. Juniors could take the history, English, solid geometry, or chemistry of the senior year and apparently do as well as the present seniors. If the seniors were to take such subjects as advanced algebra, history, physics, etc., now being taken by the juniors, assuming that they had not had the subject before, they would

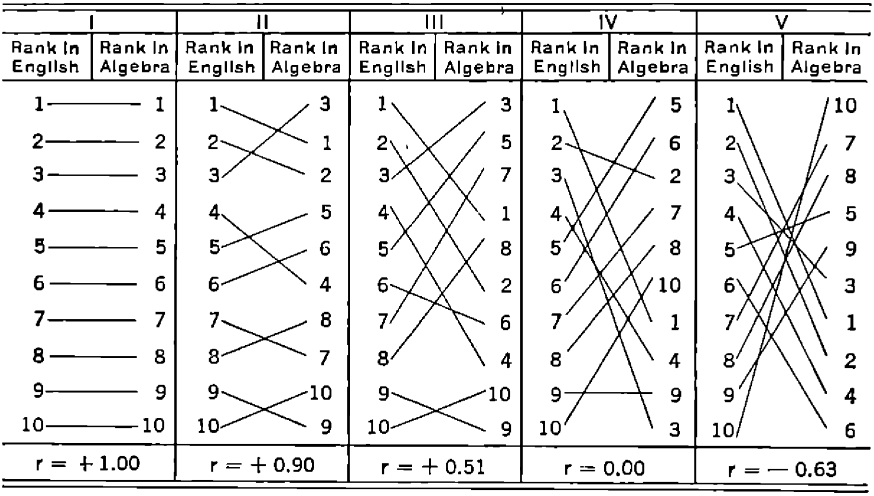
make no better showing on the average than the present juniors. Where there are two sections in the same subject in any class, the teacher's burden would be reduced and the pupils' progress increased if they could be separated into two groups as homogeneous as possible and with little overlapping.

V. Correlation

The teacher will find that in her daily work with tests a knowledge of and skill in finding the mean and standard deviation will be a sufficient kit of statistical tools with which to handle the results of testing. But the theory of testing requires a third concept to supplement central tendency and variability, namely, *correlation*. The concept of correlation and an accurate measure of correlation is necessary to validate all of our tests and to inform us as to their reliability. An accurate measure of the degree of correlation is needed to determine precisely the predictive value of a test. Correlation is needed in the critical study of many of the techniques of school administration. The unit by which correlation is commonly measured is known as the *coefficient of correlation* and the computing procedure usually used yields what is known as the "product-moment" or the "Pearson coefficient of correlation." The concept of correlation can best be illustrated by describing the Pearson coefficient of correlation, and expounding its significance. The coefficient of correlation, which is symbolized by r, runs from -1.00 to $+1.00$. -1.00 correlation signifies perfect negative correlation or perfect compensation; .00 correlation signifies a chance relationship between the two variables; while $+1.00$ signifies perfect correlation or correspondence between the two variables. These statements may be clarified by a concrete example, adapted from Gates.[1] Let there be ten pupils, 1, 2, 3, etc. Let us place them in order of their achievement in a school subject such as English so that

[1] Gates, A. I. — *Psychology for Students of Education*, p. 418, 1924.

the pupil with highest achievement in English is placed first, the pupil with next highest achievement in English is placed second, and the pupil with lowest achievement is placed last. In the second column let the order of the numbers indicate the rank in achievement of the same pupils in algebra. Lines have been drawn connecting the numbers which represent each individual. The slope of these lines indicates the displacement in position, and failure to correlate perfectly. In column I each



Illustrating the significance of coefficients of correlation (r) of various magnitudes. (Adapted from Gates)

pupil has the same rank in English that he has in algebra. The relationship is perfect and the coefficient of correlation is + 1.00. In column II there is some shifting of position. Pupil 1, who leads the class in English, is in second position in algebra. Pupil 3 leads the class in algebra. There is some crossing of lines, but the displacement is at most only one or two positions and the coefficient of correlation is + 0.90. In column III there is still more displacement. Pupil 1 is still above average in algebra, and pupils 9 and 10 are still below average in algebra. The

coefficient of correlation is +0.51. In column IV the rank in algebra is a chance order. It is as though tablets with the numbers 1 to 10 were placed in a box, shaken up, and then withdrawn one at a time. Any number might come out first; 5 actually does stand first. A purely random order like this signifies zero correlation. In column V the tendency is reversed. Pupils who are above average in English tend to be below average in algebra and vice versa. This reversal of position is called a negative correlation. In this particular instance $r = -0.63$. A sixth figure might be drawn to represent −1.00 correlation. In such a figure the rank in algebra would be exactly the reverse of the rank in English. The lines connecting each individual would spread out fanwise and would intersect in a point. Negative correlations never occur (except in rare occurrences) between two school subjects. The correlation between school subjects may be small but is invariably positive. Perfect correlation (+1.00) is also never found between two school subjects. +0.50 is a typical figure for the correlation found in practice between marks in two subjects as English and algebra. It seldom runs higher than this, but may be lower under special circumstances.

Correlation is usually exhibited by placing two distributions of scores on the same plot. This is known as a "scatter diagram." The following diagrams which are given to illustrate the coefficients of correlation of different sizes are taken from an early work by Darbishire.[1] An explanation of how the tables were derived, although the method was purely theoretical, helps make clear the meaning of correlation. To derive the table illustrating $r = 0.50$, twelve dice were rolled, six of the dice being red and six being white. The number of dice showing faces with four or more pips

[1] Darbishire, A. O. — "Some Tables for Illustrating Statistical Correlation," *Memoirs and Proceedings of the Manchester Literary and Philosophical Society, Vol. 5*, 1907.

TABLE 68

Illustrating a Correlation of 0.00

	0	1	2	3	4	5	6	7	8	9	10	11	12	
12														
11														
10					2		1	2						5
9		1		1	2	9	7	6	6		1			33
8				1	7	15	19	12	6	6				66
7			1	5	7	16	22	15	13	6	1		1	87
6			1	6	15	25	24	28	15	6	2	1		123
5			3	5	13	26	14	14	12	6	1	1		95
4			4	4	7	9	6	12	5	5				52
3				1	4	7	8	5	4	1	1			31
2				1		4				1				6
1						1			1					2
0														
		1	9	24	57	112	101	94	62	31	6	2	1	500

TABLE 69

Illustrating a Correlation of 0.25

	0	1	2	3	4	5	6	7	8	9	10	11	12	
12														
11								1						1
10						1	1	1	4	3				10
9					4	3	9	6	6	2				30
8				1	5	6	10	14	8	7	2	1		54
7			1	4	9	17	18	24	16	5	3			97
6		1		10	16	17	23	28	22	5	1			123
5			4	3	12	15	23	22	9	3	1			92
4			1	8	6	21	16	6	6					64
3		1	1	5	2	2	4	5						20
2							6	1						7
1					1		1							2
0														
		2	7	31	55	82	111	108	71	25	7	1		500

TABLE 70

ILLUSTRATING A CORRELATION OF 0.50

	0	1	2	3	4	5	6	7	8	9	10	11	12	
12														
11									1					1
10								2	1	2				5
9						3	5	6	9	5	2			30
8					2	7	13	22	14	5	3			66
7				2	2	13	16	27	12	4	2			78
6			1	5	14	25	24	24	17	4	3			117
5			2	5	17	24	19	25	11	2				105
4			5	9	8	11	16	7	6	1				63
3			2	3	5	6	2	6						24
2			1		2	3	2							8
1			1	1	1									3
0														
			12	25	51	92	97	119	71	23	10			500

TABLE 71

ILLUSTRATING A CORRELATION OF 0.75

	0	1	2	3	4	5	6	7	8	9	10	11	12	
12														
11										1		1		2
10							1		4	3	2			10
9						3	6	15	9	7	1			41
8						7	28	16	11	5				67
7				4	17	26	30	18	7					102
6			1	10	18	34	26	10	1					100
5			6	17	30	32	13	1						99
4		1	8	18	19	5	1							52
3			5	7	3	1								16
2		2	5	1	1									9
1		1	1											2
0														
		4	26	57	88	108	105	60	32	16	3	1		500

TABLE 72

ILLUSTRATING A CORRELATION OF 1.00

	0	1	2	3	4	5	6	7	8	9	10	11	12	
12														
11												1		1
10											5			5
9										30				30
8									64					64
7								81						81
6							117							117
5						104								104
4					63									63
3				24										24
2			8											8
1		3												3
0														
		3	8	24	63	104	117	81	64	30	5	1		500

was counted and then the white dice were gathered up and thrown again, and again the number of dice showing faces with four or more pips was counted. This gives a pair of numbers. This process was repeated a large number of times, 500 times to be exact. The result of these throws was then plotted as shown in table 70. Suppose that the first throw showed 7 faces with four or more pips and that in the second throw when the six white dice were rethrown only 3 faces showed with four or more pips. This throw was plotted as a point. The value 7 is found by running along the scale at the top of the figure. The point is in this column. Run up the column until opposite the 3 in the scale at the left. This definitely locates the square in which this throw is to be placed. The other four hundred ninety-nine pairs of throws were plotted in like manner. Figures indicate the frequency of pairs of throws that fell in each

square. The size of the correlation may be varied by varying the number of white dice that are picked up. If all are picked up, the correlation is .00 (approximately). If nine out of twelve are picked up, the correlation is 0.25. If three out of twelve are picked up, the correlation is 0.75. If none are picked up — that is, if the same value is allotted to each member of every pair — the correlation is 1.00.

One notes the characteristics of correlations of various degrees. In a correlation of + 1.00 all of the items lie along a diagonal which goes from the lower left corner to the upper right corner. In the correlation of +0.75 the items have departed from the diagonal slightly so that a narrow ellipse is formed. Take a score of 6 at the top. One notes that this does not always correspond to a score of 6 along the side, but that scores of 4, 5, 6, 7, 8, 9, 10 are possible. One could not predict from a score of 6 along the top exactly what the score would be in the other direction, but one could predict the score within the limits of 4 to 10. In the correlation of +0.50 the figure has become still wider. For any score such as 6 along the top the prediction is still poorer — the possible range of scores is greater. In the figure representing the correlation of +0.25 the figure is more rounded and the variability of any array within the plot is greater. By array is meant the distribution of scores for one variable for a single score in the other variable. The correlation representing 0.00 is practically round. No relationship is shown. For any one value, such as 6 along the top, there is a possibility of *any* value in the other variable. Given a value in one variable one can make no prediction of the value in the other variable. As one passes to negative correlation the elliptical shape forms along the other diagonal until −1.00 correlation is represented by a line along the diagonal running from upper left to lower right. These figures help one visualize the meaning of correlation as it passes from the circular, pure chance relationship of 0.00 through

various shaped ellipses to the thin diagonal representing perfect correlation + 1.00.

Until one is accustomed to the coefficient of correlation one does not know when it is high or low. The following table gives typical coefficients which may help one to get his bearings.

$r =$	0.90	the correlation of form *A* with form *B* of a forty-minute intelligence test in high school.
$r =$	0.80	the correlation of form *A* with form *B* of a forty-minute achievement test in high school.
$r =$	0.70	the correlation of first semester marks with second-semester marks of the same class and same teacher in a school that has a fairly good system of marking.
$r =$	0.60	the correlation of an average of elementary school marks with an average of first-year high school marks.
$r =$	0.50	the correlation of marks in one academic subject in high school with marks in another academic subject.
$r =$	0.40	the correlation of marks in an academic subject with marks on an intelligence test.
$r =$	− 0.30	the correlation of marks in an academic subject in high school with chronological age (this correlation is negative).

Correlations lower than this are so low as to be worthless in general.

REFERENCES

BROWN, W. M., and THOMPSON, G. H. — *The Essentials of Mental Measurement.* Cambridge University Press, 216 pp., 1921.

DARBISHIRE, A. O. — "Some Tables for Illustrating Statistical Correlation," *Memoirs and Proceedings of the Manchester Literary and Philosophical Society, Vol. 5,* 1907.

GARRETT, H. E. — *Statistics in Psychology and Education,* ch. 4. Longmans, Green & Company, New York; 317 pp., 1926.

GATES, A. I. — *Psychology for Students of Education,* ch. 17. The Macmillan Company, New York, 1924.

HOLZINGER, K. J. — "A Combination Form for Calculating the Correlation Coefficient and Ratios," *Quarterly Publication of the American Statistical Association,* 18: 623–627, March, 1923.

JONES, D. C. — *A First Course in Statistics*, Part 1, chs. 2–5. G. Bell and Sons, London; 286 pp., 1921.

KELLEY, T. L. — *Statistical Method*, ch. 8. The Macmillan Company, New York, 1923.

McCALL, W. A. — *How to Measure in Education*. The Macmillan Company, New York, 1922. *Ref.* Ch. 3.

OTIS, A. S. — "The Otis Correlation Chart," *Journal of Educational Research*, 8 : 440–448, Dec., 1923.

— *Statistical Method in Educational Measurements*. World Book Company, Yonkers, New York, 337 pp., 1925.

RUGG, H. O. — *Statistical Methods Applied to Education*. Houghton Mifflin Company, Boston, 410 pp., 1917.

SYMONDS, P. M. — "Variations of the Product-Moment (Pearson) Coefficient of Correlation," *Journal of Educational Psychology*, 17 : 458–469, Oct., 1926.

THORNDIKE, E. L. — *Mental and Social Measurements*. Bureau of Publications, Teachers College, Columbia University, New York, 277 pp., 1912.

THURSTONE, L. L. — "A Data Sheet for the Pearson Correlation Coefficient," *Journal of Educational Research*, 6 : 49–56, June, 1922.

— *The Fundamentals of Statistics*. The Macmillan Company, New York, 237 pp., 1925.

TOOPS, H. A. — "Eliminating the Pitfalls in Solving Correlation; a Printed Correlation Form," *Journal of Experimental Psychology*, 4 : 434–446, Dec., 1921.

YULE, G. U. — *An Introduction to the Theory of Statistics*. C. Griffin & Company, London, 381 pp., 1912.

CHAPTER XIII

NORMS AND STANDARDS

Norms

In the chapter on intelligence tests tables of norms were given but not explained. This chapter will be devoted to a discussion of those tables, — their meaning, their varieties, how they are obtained, and some of the cautions needed in using and interpreting them.

With the introduction of standardized tests and their objective scores we have a welcome new point of view in *norms* and *standards*. A *norm* is an *average; a standard* is a *goal*. Take any group of children — a freshman class of 150 in high school, to be specific. Give this class the "Terman Group Test of Mental Ability." Score the test and find the arithmetic mean of the scores. Let the mean score be 110. This score is a *norm*. Ordinarily this norm would be quite unacceptable. One expects a norm to be found from many more than 150 cases. The norms for the "Terman Group Test of Mental Ability" were found on 4,000 to 10,000 children in each grade. Also one expects a norm to be found from a wider selection of children than one school contains. One expects many schools to furnish scores to be used, selected from many different cities and towns in many different states.

There are several types of norms. There are *grade* norms which are averages of scores of children by grade in school. For instance, an 8th-grade norm is an average of scores made by children in the 8th grade. Since any one room is quite apt to be far from the average of other rooms, an 8th-grade norm ought to be

the average of many 8th-grade classes taken from schools in various localities. The following set of averages on a spelling test given in thirteen 4th-grade rooms in a city school system illustrates variations in achievement.

TABLE 73

AVERAGES ON A SPELLING TEST GIVEN IN THIRTEEN FOURTH-GRADE ROOMS IN A CITY SCHOOL SYSTEM

ROOM	AVERAGE SCORE
A	12.7
B	13.8
C	14.2
D	20.4
E	20.5
F	20.9
G	23.4
H	24.4
I	25.2
J	26.2
K	27.6
L	30.3
M	32.2

It is plain that the averages obtained from room A or room M could not be used with justice as norms. A much fairer norm for this city would be the average of all thirteen rooms, which is 22.5. Yet even this latter figure might not fit conditions in another city. We expect norms to be widely representative.

Age norms are averages of the scores on a test by age. The average of all 12 year olds would constitute the 12-year age norm. The same cautions are again necessary — the group whose scores are to be averaged should be a large group and a representative group. In obtaining age norms one should be careful not to leave out children who are greatly accelerated or retarded in school.

For instance, in getting 12-year age norms in the "Otis Advanced Examination," Otis found some children in grade 4 and some in the second year of high school who were 12 years old. These, of course, should be included and without them the average might have been too high or too low.

This leads to the fact that it is very difficult (if not impossible) to obtain true age norms in the high school age levels. By the first year of high school one half the boys and girls of the appropriate age have dropped out of school. Not all, but a large majority of those who dropped were those who were unsuccessful in school and who, naturally, if they could be tested, would test low and would lower the averages if their scores were included. Accordingly all age norms of 14 years and over are to be suspected as being too high — just how much too high cannot be stated with accuracy. This phenomenon is known as *selection.* Selection makes the determination of age norms in the high school next to impossible — hence we shall make very little use of age norms in discussing high school tests. This does not mean that age norms are not desirable. As a matter of fact, age norms theoretically are much more serviceable than grade norms. A grade norm is not rigorously defined — grades differ according to courses of study. But an age is something more or less objective and age norms are consequently more stable than grade norms. If someone is able to give an intelligence test to all the 16 year olds of a community, thereby getting an age norm, he will be doing scientific education a great service.

In Chapter IV one will find for the Terman and Otis tests *mental age norms.* These norms are scores equivalent to the mental ages found on the Stanford-Binet. Both the Terman test and the Stanford-Binet (individual) test have been given to the same group of children. By statistical procedure the mental ages of the individual test are transmuted into the scores of the group test and the tables in Chapter IV show the equivalent

scores. Mental age norms have several advantages. They give an easy and justifiable method of translating scores found on the group test into ages, which is indeed necessary if one is to find I Q's, etc. Another advantage of mental age norms is that one feels certain that the norms are as representative as the Stanford-Binet scale itself. One can obtain mental age norms in a single school of 300 and be confident that although the school may represent a selected group of children, the norms are representative norms. This is so because the scores on the test are turned into Binet mental ages which have already become "standardized."

There are also cautions to be observed in using mental age norms. In the first place, in selecting tests for the Stanford-Binet scale in the average adult level Terman did not make sure that these tests exactly fit average adults. As a matter of fact, it is now known that the average adult tests in the Stanford-Binet are too hard for average adults. What Terman did was to make sure that they were more difficult than the 14-year-old tests and less difficult than the superior adult tests. Terman was not better able to test unselected 16 year olds than the users of group tests. But once the tests were placed in the upper levels of the Stanford-Binet they became standards — artificial standards — and mental ages obtained from the upper levels of the Stanford-Binet have since been used exactly as though they were actual age norms. In using mental age norms on a group test one should note that these are not necessarily the same as actual age norms obtained from averages.

Another caution to be noted is that no two mental tests are equivalent — they do not measure exactly the same thing — the same kind of intelligence. So it is not perfectly justifiable to say that such and such a score on the Otis test is equivalent to such and such a mental age on the Stanford-Binet. Everyone can see that it is not justifiable to say that the wealth of Henry Ford is

equivalent to the highest intelligence, because wealth and intelligence do not always go together. Likewise, in a *very slight way* two intelligence tests do not go together and hence it is not exactly fair to make a comparison of their scores. The man who makes the highest score on the Otis test may not make the highest mental age on the Stanford-Binet. In reading the equivalent mental age from a given score in the mental-age norms, however, the highest Otis score corresponds to the highest mental age.

If one must turn scores into ages, use mental-age norms rather than actual-age norms — we know exactly what mental-age norms are, we do not know what age norms obtained from averages on the high school level are. In using mental-age norms remember their limitations. The real conclusion is: *Do not turn test scores into ages for use in high school. We shall later see that we can do everything that we want with the tests by using the scores themselves*, and comparing them with grade norms when necessary.

A third type of norm is the *percentile norm.* These are given in Chapter IV for the Otis, Terman, and Miller tests. Percentile norms may be either grade norms or age norms, — in Chapter IV they are all grade norms. The tables of percentile norms run in two directions — one way there are grades and the other way per cents. The scores in any place in the table tell *what per cent of a grade equals or exceeds the score given.* For instance, in the norms given for the "Terman Group Test of Intelligence" look at the column for grade 7. Opposite 50% is the score 68. This means that 50%, or half of the children of the 7th grade, equal or exceed a score of 68 on this test. Such a score equaled or exceeded by 50% is called the *median.* The score of 88 in the same column signifies that 25% of the 7th-grade children reach or exceed this score. Such a score reached or exceeded by 25% of a group is called the *upper quartile.* To find what score is reached or exceeded by 75% of the group (*lower quartile*) run the eye across until the score 51 is reached.

A fourth kind of norm, known as the "sigma index," developed by Franzen,[1] is really the most scientific of all types of norms yet proposed. As none of the high school tests is yet supplied with norms of this type, an extended description is not included here. The "sigma index" is the distance of any score on a test from the mean score for any age group in terms of the standard deviation of that age group.

The question of interpretation of a score on a test is often raised. Suppose a boy in the 9th grade (first-year high school) makes a score of 140 on the "Terman Group Test of Mental Ability." Turning to the grade norms for the Terman Test and looking in the 9th-grade column, the nearest score one finds is 142, which corresponds to a percentile of 15. We may say, then, that this boy made a score equaled or exceeded by only slightly more than 15% of high school freshmen in general. Looking along the line of the 50 percentile or median, one finds that the average high school junior makes a score of 138 and the average high school senior a score of 147. So our boy who makes a score of 140 is equivalent in intelligence to the average high school junior or senior. The same sort of interpretation may be given a class average as well as an individual score.

From this it is seen that we no longer interpret test results in terms of *perfection* or *passing* but in terms of *averages*. A single test score is interpreted by comparing it with other scores in the same group by means of percentiles or with the average performance of different groups. The aim of teaching is no longer to raise as many as possible to perfection or to prevent as many as possible from failing, but *to raise the average of the class*. The most significant fact that we can give about any class is its average.

[1] FRANZEN, R. H., and HANLON, W. H. — *The Program of Measurement in Contra Costa County*, pp. 5–14, 1924.

Standards

A standard is a goal. Usually a goal is not the same as a norm. One usually expects the goal of instruction to be better than the average performance. On the other hand, the norm is only too often taken as the standard and there is much confusion in most people's minds regarding these two concepts. It may be stated as an axiom that *standards cannot be determined from the results of testing alone,* although the results of testing may give much aid in setting standards. Setting standards from the results of testing alone implies that what is should be. This no educator would accept. Standards must depend, in the first place, on objectives, or aims, and on relative values.

To make the point clearer let us state an actual case. McCall [1] has determined that at the end of the eighth grade pupils are able to make an average score of 60.9 on the Thorndike-McCall reading test (this is a score, not a per cent). Shall this score be set up as a standard or goal for the end of the 8th grade? Shall teachers aim to bring their classes up to this level and no higher in reading? Shall children who do not reach this level be required to take more instruction in reading until they are able to read as well as a score of 60.9 signifies? That is, shall what children on the average accomplish be held as a standard or goal of accomplishment? It must be evident that if the norm is to be made a standard, approximately 50% will fail to meet the standard. Before these questions are answered we shall have to answer such questions as: Is reading ability equivalent to a score of 60.9 good enough to guarantee understanding of the reading in history, geography, and civics required in the 8th grade? Is it good enough to insure ability to do the work required in the 9th grade? Is enough emphasis given to reading ability in the 8th grade as compared with other subjects? Can the level of reading be raised by improved methods

[1] McCall, W. A. — *Directions for Using Thorndike-McCall Reading Scale,* p. 4.

of instruction? What level of reading ability is required in various occupations, for general culture and pleasure, for effective citizenship, etc.? Only until these and similar questions are answered will we be in a position to set standards. It is certainly helpful to know quantitatively how good reading ability now is in the 8th grade, for any standards set up should be in terms of more or less than present ability, but one must not confuse present performance with desirable performance.

One must confess that education has at the present time set few standards in terms of test scores. Standards, when set, have been verbal or logical. There is a great need for quantitative standards. One reason for the failure to set adequate standards is the confusion that exists between norm and standard and the widespread use of norms as standards. Courtis [1] recognized the difference between norm and standard in his work on arithmetic tests and proposed standards which were not norms. With reference to standards which are set in the "Standard Practice Tests in Arithmetic," Courtis [1] says,

> The standards above have been set after a careful investigation of the actual speed of work of children in the grades, and are believed to represent the speeds at which children can work without strain.

Thorndike sets up standards for algebraic computation:

> We recommend that practice on such fundamentals as appear in Nos. 1 to 11 should be sufficient to secure the ability to do such tasks (without full checking) in 30 seconds or less with at least nine right answers out of ten; with Nos. 12 to 17 the time may be increased to 40 seconds. The pupil should be able, if time for full checking is available, to turn in such work substantially errorless.
>
> Our standards of precision may be thought to be too severe for ordinary attainment since these gifted minds do not far surpass it, except after warning to be careful. However, there should be no very great superiority of

[1] Courtis, S. A. — *Teacher's Manual*, Courtis Standard Practice Tests in Arithmetic, p. 8. Copyright 1920 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

gifted over ordinary minds in the surety of these routine computations or in the clerical work of reading and copying. Having nearly four times as long, the school pupil ought to do nearly as well. Relaxation to seventeen out of twenty right is defensible, but not in our opinion desirable.[1]

Courtis and Thorndike, keeping in mind the aims of the instruction, set these two standards by sagacious scrutiny of actual performance. Not much better than this can be done, although it is generally safer to obtain the concensus of opinion of several experts in order to eliminate personal bias.

The best example of an objective method of determining standards is found in the subject of handwriting. Ayres, Ashbaugh, Koos,[2] and Kirk [3] have studied the handwriting of men in clerical positions and the handwriting ordinarily used in correspondence. Koos studied 1,053 specimens of handwriting taken from social correspondence and 1,127 specimens written by employees in various occupations. He also secured the judgment of 826 adults as to what they considered adequate and inadequate handwriting for social correspondence. He concludes:

It is difficult to see why, for the use under consideration, a pupil should be required to spend the time necessary to learn to write better than the quality 60. The quality 60 on the Ayres Measuring Scale for Adult Handwriting, which we have set up as the ultimate standard of attainment for all school children for purely social purposes, is adequate for the needs of most vocations. This will apply to labor, skilled and unskilled, as well as to the professions, exclusive of teaching in the elementary schools. For that large group who will go into commercial work, for telegraphers, and for teachers in the elementary schools, it will be necessary to insist upon the attainment of a somewhat higher quality, but hardly in excess of the quality 70.

The writer has determined standards of achievement of first-year algebra in terms of the Hotz algebra scales by the round-

[1] THORNDIKE, E. L. — *Psychology of Algebra*, pp. 363–364, 1923.

[2] KOOS, L. V. — "The Determination of Ultimate Standards of Quality in Handwriting for the Public Schools," *Elementary School Journal*, 18: 422, Feb., 1918.

[3] KIRK, J. G. — "Handwriting Survey to Determine Grade Standards," *Journal of Educational Research*, 13: 181–188, 259–272, March, April, 1926.

about method of first determining the intelligence necessary to be successful as an engineer, an occupation that probably demands as much use of algebra as any, and then translating this intelligence into ability in algebra on the Hotz scale.[1]

From these illustrations it will be seen that the determining of standards is by no means the easy process that the determination of norms is. I will sum up this section by stating three guiding principles in the determination of standards. First, standards must take into account relative values between subjects or activities. The child's whole daily life must be considered. No standard should be set so high that time is robbed from other important subjects or activities. Time should not be required in attaining a standard that will diminish the time for desirable athletic and social activities, recreation, and home life. In setting standards one must have consideration for child growth and development and must not set standards which require undue pressure or drill for accomplishment. Here norms will serve as safeguards in setting standards. One should hesitate in setting a standard much above a norm unless he is sure that the standard can be attained without strain, or unless allowance is made for reducing standards in other subjects. Norms represent progress made under existing conditions and allotment of time. If new developments in teaching methods are used, or if motivation which does not mean undue expenditure of energy or effort is employed, then standards may be safely raised without consideration of the effect of such a raise on other subjects or activities.

Secondly, standards should be set with due consideration of other school requirements both at the same time or in the future. Standards of reading should be adequate to guarantee ability to do the work in other school subjects using reading (and they all do use reading). Standards in arithmetic should be sufficient to

[1] Symonds, P. M., — *Special Disability in Algebra.*

satisfy the demands of algebra (if there is a chance that algebra will be studied) or the demands of business training. Standards in algebra should be sufficient to meet future demands in trigonometry and higher mathematics, or in physics, optics, mechanics, etc. Standards in French and German should be sufficient to meet demands for reading scientific or literary material in these languages if such demands are in prospect. Standards in English composition should be adequate for the demands of the school in other subjects.

Thirdly, standards should be adequate to satisfy demands of life outside school. The examples given for handwriting illustrate this. Handwriting should satisfy the needs of business and general correspondence and no more than these demands. In general it was found that the standards in handwriting set by schools were higher than those needed in business or in private life. So this third criterion may lower existing standards as well as raise them.

Before closing this topic of standards it ought to be said that in the final analysis standards must be set up for individuals as well as for groups. It is well enough to say that *every one* ought to be able to spell, write, read, and compute, according to such and such a standard, but when we come to apply this in practice, many exceptions will occur. There are some unfortunate individuals who can never learn to read according to any standard we may establish. These children are feeble minded or very low grade mentally — for them even minimum standards must be lowered. There are other children for whom it would be waste to say "You need to achieve only such and such a standard." Some children ought to be encouraged, indeed required, to achieve much more than the minimum, or typical, standard. These are the bright or gifted children in our schools. Not only this, but if we try to apply standards by age as well as grade, we find that there are great individual differences. Where such

achievement correlates highly with intelligence, we can first determine the intelligence of the child and then translate this intelligence into standards of achievement. The Accomplishment Ratio, a unit developed by Franzen,[1] and Monroe and Buckingham,[2] is a unit which is designed to aid in setting standards in terms of individual achievement. According to this method each child is set a goal or standard each year in proportion to his mental ability, — those with high mental ability are expected to achieve a higher standard than those with low mental ability.

Causes for Variations in Norms

Workers in educational measurements have been much disturbed because norms obtained by different workers for the same test may vary considerably. A good example of this occurred in the standardization of the "National Intelligence Test." The standardization was made on the basis of testing carried on in schools of Washington, D. C., and of Pittsburgh and the norms were the averages of achievement on the tests in those cities. It was somewhat disconcerting to find later that these norms were considerably too high, especially the age norms. Terman and Whitmire[3] corrected these norms by additional testing at Vallejo, California. Even Terman's norms were too high, as subsequent testing in other localities showed. It was not until the publishers received returns from thousands of cases in many parts of the country that really stable and satisfactory norms were available. Let us see what are some of the factors that cause norms to fluctuate. This is not the place to set forth

[1] Franzen, R. — "The Accomplishment Quotients: A School Mark in Term of Individual Capacity," *Teachers College Record*, 21: 432–440, Nov., 1920.

[2] Monroe, W. S., and Buckingham, B. R. — *Teachers Handbook*, Illinois Examination, p. 31, 1920.

[3] Terman, L. M., and Whitmire, E. D. — "Age and Grade Norms for the National Intelligence Scales *A* and *B*," *Journal of Educational Research*, 3: 124–131, Feb., 1921.

evidence which demonstrates that these are the factors which influence norms, but illustrations will be given which will make concrete the existence of these influences.

VARIATIONS DUE TO DIFFERENCES IN LOCALITY

It is known that locality has much to do with differences in norms. Frost [1] has made a study of comparative achievement of country and town schools. The following table summarizes his findings for 12-year olds.

TABLE 74

AVERAGE SCORES MADE BY 12-YEAR OLDS IN CITY AND COUNTRY SCHOOLS (FROM FROST)

TEST	LOUISVILLE, KY.	MADISON CO., KY.	
		Nine Mo. School	Six Mo. School
Trabue Language Scale B	14.5	11.1	9.0
Courtis Arithmetic B			
Addition	6.7	3.3	1.0
Subtraction	7.9	4.5	1.0
Multiplication	7.1	3.1	0.9
Division	7.6	2.3	0.8
Thorndike Reading Alpha	6.5	5.9	5.1

Other city and country schools were included, but the table presented above is sufficient to show the superiority of city-school achievement over rural-school achievement. It is now a well-established fact that city-school achievement is superior to rural-school achievement and must be allowed for in interpreting norms. Most norms are obtained from testing carried on in city schools. Consequently test results obtained in rural schools ought not to be compared with norms obtained in city schools. Probably

[1] FROST, N. — *A Comparative Study of Achievement in Country and Town Schools.*

we should have norms for both city and country schools. This ought not to be interpreted as meaning that different standards should be held for city and country — perhaps the standards should be the same or perhaps sociological considerations would show that in some subjects standards should be different between city and country.

The testing during the war permitted comparisons between states on the basis of scores on "Army Alpha." Alexander[1] has

TABLE 75

MEDIAN SCORES MADE BY WHITE RECRUITS IN "ARMY ALPHA" IN FORTY-ONE STATES (FROM ALEXANDER)

STATE	MEDIAN ALPHA SCORE	STATE	MEDIAN ALPHA SCORE
Oregon	79.9	New Hampshire	61.9
Washington	79.2	Missouri	59.5
California	78.1	South Dakota	58.3
Connecticut	73.6	North Dakota	57.1
Idaho	73.5	Wisconsin	56.5
Utah	72.2	Virginia	56.3
Massachusetts	71.6	Maryland	56.2
Colorado	69.7	Indiana	56.1
Montana	68.5	Oklahoma	52.5
Vermont	67.5	Texas	50.9
Ohio	67.3	New Jersey	48.7
Maine	67.0	South Carolina	47.4
Nebraska	66.2	Tennessee	47.2
Pennsylvania	65.1	Alabama	46.3
New York	64.5	Louisiana	45.2
Iowa	64.4	North Carolina	43.2
Minnesota	64.0	Georgia	42.2
Kansas	63.9	Arkansas	41.6
Illinois	63.8	Kentucky	41.5
Michigan	63.3	Mississippi	41.2
Rhode Island	62.9		

[1] ALEXANDER, H. B. — "A Comparison of the Rank of American States in Army

computed the averages (norms) for each state and has shown that there are startling differences between states. The facts are reproduced in Table 75.

From these figures it is seen that one western state makes an average score on "Army Alpha" (80) almost twice as high as one southern state (41). Such differences are startling and unexpected. They illustrate forcibly the fact that locality has much influence on norms. An age norm obtained from testing in Washington, California, or Oregon is certainly too high for the country as a whole. (Practically all tests emanating from Stanford University have been standardized on California children.) An age norm obtained from Kentucky or Mississippi is almost sure to be too low. Perhaps each state should have its own norms. At least such considerations should warn one against making invidious comparisons in any one school, using norms which are not comparable. Norms should be scrutinized carefully before they are used for comparative purposes.

Franzen and Hanlon [1] have tested all the schools in Contra Costa County, California. In the fifth grade an average measure of ability ranged all the way from 54 for one school to 114 for another school; in the sixth grade the scores ranged from 44 to 141; in the seventh grade the scores ranged from 30 to 116; and in the eighth grade from 75 to 138. These were average scores for a school. The figures demonstrate that communities vary widely in ability of the children in school.

VARIATIONS DUE TO DIFFERENCES IN RACE

The testing in the army also showed that in testing different racial or national groups differences in the averages are obtained. On a combined scale, using the results on both Army Alpha and Army Beta, the following means are found (taken from Brigham's

Alpha and in Social Economic Status," *School and Society*, 16: 388–392, Sept. 30, 1922.

[1] *Op. cit.*

A Study of American Intelligence [1]) for foreign-born white draft by country of birth.

TABLE 76

MEAN SCORE IN INTELLIGENCE ON A COMBINED SCALE DERIVED FROM ARMY ALPHA AND ARMY BETA OF FOREIGN-BORN WHITE RECRUITS (FROM BRIGHAM)

	No. Cases	Mean
England	411	14.9
Scotland	146	14.3
Holland	140	14.3
Germany	308	13.9
Denmark	325	13.7
Canada	972	13.7
Sweden	691	13.3
Norway	611	13.0
Belgium	129	12.8
Ireland	658	12.3
Austria	301	12.3
Turkey	423	12.0
Greece	572	11.9
Russia	2340	11.3
Italy	4009	11.0
Poland	382	10.7
U. S. (colored)		10.4

It is not asserted here that these differences are due to differences in native ability. They probably are to some extent. They are probably also due to differences in familiarity with the English language and other factors. What is emphasized here is that *there are differences* in the results of testing different racial groups. Accordingly one should know what the racial composition is of the group used in obtaining norms. If the norms are obtained mainly from a group of native stock the norms cannot be applied indiscriminately in comparing groups of other racial composition.

[1] Pp. 120–121.

VARIATIONS DUE TO DIFFERENCES IN SEX

The two sexes usually show no appreciable difference in standardized testing, but in some subjects they do. Van Wagenen [1] has shown that there is a consistent sex difference arising in the giving of history tests. The following table gives Van Wagenen's norms on his history test showing clearly the sex differences.

TABLE 77

VAN WAGENEN AMERICAN HISTORY SCALES: GRADE MEDIANS

Grade	Information A		Thought A		Character A	
	Boys	Girls	Boys	Girls	Boys	Girls
4	5.6	4.4	2.8	2.3	2.0	1.9
5	10.0	7.6	6.4	4.1	4.4	4.0
6	13.6	10.3	12.9	8.1	7.2	6.4
7	15.2	10.4	16.1	12.3	10.3	9.3
8	17.7	13.9	19.9	15.3	12.6	11.5

Accordingly one should ask in using history norms whether they were obtained from testing boys or girls or both, and if both, in what proportions. Also if one is using norms to compare results in testing a group of boys or girls, corresponding allowances should be made.

VARIATIONS DUE TO DIFFERENCES IN INCENTIVE

That differences in incentive to do school work make a difference in achievement is self-evident. It has been shown many times that increased interest in school work, increased motivation, however produced, makes marked differences in the results on standardized tests. This refers particularly to subject matter or achievement tests. The classic experiment in which this is demonstrated is

[1] VAN WAGENEN, M. J. — *Historical Information and Judgment in Pupils in the Elementary School.*

that carried out by Franzen in Garden City, Long Island. Franzen,[1] by various devices which promote interest and increase motivation, such as regrading, showing the results of standardized tests to children, and the like, was able to make distinct advances in test averages.

Accordingly one should know how much motivation there is in a class in which norms are obtained. Sometimes this enthusiasm which emanates from some able administrator influences the results in a whole school system. Of course this enthusiasm or motivation is difficult to gauge directly and perhaps cannot be allowed for in testing to obtain norms or in using norms for comparative purposes.

VARIATIONS DUE TO METHODS OF TEACHING

Methods of teaching make immediate and effective differences in achievement as measured by standardized tests. Many studies show this. A typical study, O'Brien's *Silent Reading*,[2] shows the gain both in comprehension and rate of silent reading by intensive drill in acquiring speed in reading.

TABLE 78

SPEED OF READING IN WORDS PER MINUTE ON COURTIS SILENT READING TEST (FROM O'BRIEN)

GRADE	NO. PUPILS	BEFORE PRACTICE	AFTER PRACTICE
IV	236	155.7	236.4
V	154	190.7	277.8
VI	128	197.8	292.6
VII	206	205.6	321.6
VIII	92	220.8	393.0

[1] FRANZEN, R. H. — *The Accomplishment Ratio.*

[2] O'BRIEN, J. A. — *Silent Reading with Special Reference to Methods for Developing Speed, 1921.*

TABLE 79

COMPREHENSION OF READING AS DETERMINED BY NUMBER OF QUESTIONS CORRECTLY ANSWERED IN COURTIS SILENT READING TEST

GRADE	NO. PUPILS	BEFORE PRACTICE	AFTER PRACTICE
IV	97	23.9	34.8
V	47	29.6	42.1
VI	57	33.7	48.4
VII	50	45.3	58.0
VIII	46	47.1	57.8

These differences, as far as their possibility of affecting norms goes, might be called serious. If such differences can be made after a few months of specialized drill or training, one questions the validity of all our norms. There are few teachers to-day who so understand the laws of learning and their application to actual teaching situations that they can produce such differences as exhibited above. Even so, one feels unsafe in using norms when the kind of teaching or method is undefined.

Factors such as order of topics in a course may play havoc with norms, particularly with high school subjects. Mere change in a textbook may throw the achievements of a class out of line from norms determined in classes using other textbooks or following a different order of topics. For this reason high school tests can never be static, permanent instruments — they must change with every change in organization of the subject and norms must change with them. I quote from Hobbs [1] on this factor in using the Hotz Algebra tests:

Very little work in addition and subtraction, however, had been done for some weeks preceding the giving of the Addition and Subtraction test. The time had been spent in graph work and work with equations. As a

[1] HOBBS, J. B. — "Results in Giving Hotz Algebra Scale Tests," *School and Society* 12: 353, 354, Oct. 16, 1920.

result, a very large number in the class forgot, when adding fractions, to put the denominators down, thinking that they were working with equations and that therefore the common denominator disappeared as it does in equation work. Many had the work all done except for this — *i.e.*, had the numerators added correctly. As soon as I spoke to the class about it afterwards they recognized their mistake and were quite chagrined to think that they should have made it. It is evident, then, that the tests are, as might be expected, greatly influenced by the time that they are given relative to the order of taking up the different phases of the subject in class.

VARIATIONS IN NORMS DUE TO PRACTICE EFFECT

The preceding factor leads us to the present factor. *Practice effect* may be defined as the increase in score due to repeating a test, or even a second form of the same test. Much to the dismay of psychologists, it has been discovered that a second trial on an intelligence test results in scores some 5% to 10% higher than the original scores. Merely repeating a test yields a higher score. I have found in using the "National Intelligence Tests" that Scale *A* has a practice effect of 10.0 points in the score and Scale *B* of 7.0 points in the score. Gates [1] has shown that the practice effect on the Thorndike-McCall Reading tests is negligible. The practice effect appears in certain tests and not in others. I surmise that practice effect is found more in tests in which either the operation or the material is unfamiliar than in tests where the operations and material have been a matter of much practice (as reading), although this needs experimental verification.

This matter of practice effect must be allowed for. So far as possible norms should be obtained on untested groups, although this is becoming less and less possible as the use of tests becomes more widespread. If norms are obtained on groups that have become "test wise," this fact should be noted in a statement accom-

[1] Gates, A. I. — "A Study of Depth and Rate of Comprehension in Reading by Means of a Practice Experiment," *Journal of Educational Research*, 7 : 37–49, Jan., 1923.

panying the norms. In making comparisons with norms one should be careful not to compare test-wise groups or individuals with norms obtained on untested groups. Thorndike in his "Intelligence Test for High School Graduates" attempts to cut down this practice effect by allowing a group taking the test to inspect a "practice form" for ten minutes before starting work on the regular test itself. Thorndike[1] says that this practice effect falls off rapidly with each subsequent testing and is practically negligible after the third re-test.

This array of factors may tend to discredit the use of norms. Two remedies have been proposed. One is that advocated by Sandiford,[2] who suggests that instead of one set of norms we have many sets. Each locality, each race, each sex, would have its set. Norms would be carefully defined by stating the type of place they were obtained in, the methods used, the course followed, the amount of testing experience, etc. That is, one would try to get norms which fit each particular situation as closely as possible, so that the effect of instruction itself would be the only difference between the situation in which the norm was obtained, and that in which the testing is done, with which it is to be compared. The other remedy is at the other extreme. Make norms as representative as possible. Have testing carried out in different sections of the country, with different racial groups, in city and country. Have the testing for norms done in good schools and in bad schools, and in classes under good teachers and bad teachers. Be careful to see that no one text or organization of work is included, but that many different points of view with regard to work is represented. That is, make norms as far as possible include a large number of cases with all possible variations of

[1] THORNDIKE, E. L. — "Practice Effects in Intelligence Tests," *Journal of Experimental Psychology*, 5 : 101–107, April, 1922.

[2] SANDIFORD, P. — "The Standardization of Tests and Scales," *Journal of Educational Research*, 7 : 14–27, Jan., 1923.

factors which influence norms. The problem still remains of weighting the various factors.

The practical solution is one which strikes a mean between these two extremes. Factors which have a large influence and which may be rigorously defined may be taken into account in obtaining norms, other factors may be disregarded.

It is safe to say that separate norms should be obtained from schools for white children and for colored children. Separate norms should be obtained for cities and for rural districts. Subjects that show a distinct sex difference should have separate norms for the sexes. Separate norms should be given for first, second, and third repetitions of the test. As far as the factors concerning motivation or arrangement of subject matter or teaching method go, perhaps norms should endeavor to represent typical conditions as far as possible. At all events, whatever norms a test maker presents, or whatever norms a teacher uses, these factors which influence norms should be taken into account.

In the early days of testing a teacher whose class exceeded the norm was elated and a teacher whose class was below the norm felt ashamed. Some teachers with C sections have refused to test their classes with standardized tests, fearing the comparisons that would result. The day of such indiscriminate conclusions with regard to the results of testing are passing. Now when a teacher finds a class above or below the norm, she asks first, what does the norm represent — what section of the country, what race, what kind of a school, what motivation, the results of what methods of teaching and organization of subject matter, and then compares her class with the norm in the light of these factors.

Standardize is often found in the literature of measurements and should not be confused with *standard* as I have used it in this chapter. *To standardize* means to construct a test on certain well-known scientific principles, to construct directions for giving

and scoring it, to obtain representative norms for the test, and to determine how accurate and how significant a test it is. So a *standardized test* is one carefully constructed and for which definite rules for giving and scoring have been prepared, for which representative norms are available, and whose significance and accuracy is known.

REFERENCES

ALEXANDER, H. B. — "A Comparison of the Rank of American States in Army Alpha and in Social-Economic Status," *School and Society*, 16 : 388–392, Sept. 30, 1922.

BRIGHAM, C. C. — *A Study of American Intelligence.* Princeton University Press, 210 pp., 1923.

BUCKINGHAM, B. R. — "Standards in Education," *Journal of Educational Research*, 10 : 1–18, June, 1924.

COBB, M. V. — "The Limits Set to Educational Achievement by Limited Intelligence," *Journal of Educational Psychology*, 13 : 445–464, 546–555, Nov., Dec., 1922.

COURTIS, S. A. — *Teachers Manual to Accompany the Courtis Standard Practice Tests in Arithmetic.* World Book Company, New York; 63 pp., 1920.

FRANZEN, R. H. — *The Accomplishment Ratio*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 125; Bureau of Publications, Teachers College, Columbia University, New York; 59 pp., 1922.

FRANZEN, R. H., and HANLON, W. H. — *The Program of Testing in Contra Costa County*, 94 pp., 1924.

FROST, N. — *A Comparative Study of Achievement in Country and Town Schools*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 111, New York; 70 pp., 1921.

GATES, A. I. — "Study of Depth and Rate of Comprehension in Reading. *Ref.* Ch. 5.

— "The Unreliability of M. A. and I. Q., Based on Group Tests of General Mental Ability," *Journal of Applied Psychology*, 7 : 93–100, March, 1923.

HOBBS, J. B. — "Results on Giving Hotz Algebra Scale Tests," *School and Society*, 12 : 353, 354, Oct. 16, 1920.

KOOS, L. V. — "The Determination of Ultimate Standards of Quality in

Handwriting for the Public Schools," *Elementary School Journal*, 18 : 423–446, Feb., 1918.

MONROE, W. S., and BUCKINGHAM, B. R. — *Teachers Handbook*, Illinois Examination. University of Illinois, Bureau of Educational Research, p. 31, July, 1920.

MONROE, W. S. — *An Introduction to the Theory of Educational Measurements. Ref.* Ch. 3.

O'BRIEN, J. A. — *Silent Reading with Special Reference to Methods for Developing Speed.* The Macmillan Company, New York; 289 pp., 1921.

RENSHAW, S. — "The Intelligence of Teachers in Training with a Note on Practical Effects with Intelligence Tests," *Journal of Educational Research*, 7 : 28–36, Jan., 1923.

SANDIFORD, P. — "The Standardization of Tests and Scales," *Journal of Educational Research*, 7 : 14–27, Jan., 1923.

SYMONDS, P. M. — "The Accuracy of Certain Standardized Tests for School Classification," *Journal of Educational Research*, 9 : 315–330, April, 1924.

— "A Second Approximation to the Curve of the Distribution of Intelligence of the Population of the United States, with a Note on the Standardization of the Stanford Revision of the Binet-Simon Scale," *Journal of Educational Psychology*, 14 : 65–82, Feb., 1923.

— *Special Disability in Algebra*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 132, New York; 89 pp., 1922.

TERMAN, L. M., and WHITMIRE, E. D. — "Age and Grade Norms for the National Intelligence Tests Scales *A* and *B*," *Journal of Educational Research*, 3 : 124–131, Feb., 1921.

THORNDIKE, E. L. — "On the Improvement in Intelligence Scores from Fourteen to Eighteen," *Journal of Educational Psychology*, 14 : 513–516, Dec., 1923.

— "Practice Effects in Intelligence Tests," *Journal of Experimental Psychology*, 5 : 101–107, April, 1922.

THORNDIKE, E. L., and Others — *Psychology of Algebra. Ref.* Ch. 3.

VAN WAGENEN, M. J. — *Historical Information and Judgment in Pupils in the Elementary School. Ref.* Ch. 9.

YERKES, R. M., ed. — *Psychological Examining in the United States Army. Ref.* Ch. 4.

CHAPTER XIV

CRITERIA FOR THE CHOICE OF TESTS

The foregoing chapters have been sparing in their description of standardized tests. Many tests that have been published for high schools have not been mentioned, other tests (some very good ones) have been mentioned only in passing, because at the present time the measurement movement for high schools is in a period of rapid development, and tests which may be described to-day will be supplanted by better tests a few years hence. This transiency of standardized tests for high school subjects is due in part to the newness of the movement, but is something that may always be expected. High school subjects continually change as new knowledge is added, as standards of value change, and as more and more work is done in making vital the materials in secondary education. For many workers in educational measurements it is more worthwhile to be able to recognize a good test by its earmarks than to know intimately the many tests — good, mediocre, and poor — that exist to-day.

Criteria of a good test may be discussed under the following heads:

I. Validity
II. Reliability
  A. Objectivity
III. Scaling
IV. Norms
V. Cost
VI. Ease of Scoring
VII. Duplicate Forms
VIII. Standardized Directions

Choose a Test because It Is Valid

The first concern of one who wishes to choose a standardized test is to make sure that the test really measures what it purports to measure. The adequacy and detail with which a test is a measure of a trait, function, or school subject is called its validity. Validity is measured or determined by the correlation of scores on the test with some independent criterion of the school subject in question. The criterion in the case of high school subjects is some other measure of the subject which has no reference to the test whose validity is being considered.

In the case of achievement tests the independent criteria to be used for validation are few in number. One must usually fall back upon school marks or teachers' estimates of achievement as a criterion. Often where several tests have been constructed in a subject, it is possible to determine the relative validity of the separate tests by constructing the composite of all the tests, and determining the correlation of each test with this composite.

CHOICE OF MATERIAL

In judging the validity of a test one should investigate how the material of the test was assembled. Good standardized tests usually contain statements, either in the manual of directions or in articles in technical periodicals, describing how the material for the test was gathered. One should search for such a statement, and if none is to be found he may immediately suspect the care used in constructing the test. Some tests, such as the "Briggs Form Test," whose validity is evident, do not require extensive work in selection of material and hence no statement can be found. Some of the older standardized tests, such as the "Hotz Algebra Scales," would have withstood better the attacks of critics if more time had been spent in obtaining valid material and less time in scaling. If there is no description as to how the mate-

rial was assembled, one may suppose that it was hastily collected much as a teacher collects material for an informal test. With material gathered in this way a test is almost sure to be short-lived and doomed to disappear because of its inadequate sampling or because of its irregular emphasis of material.

In the case of achievement tests most of the validation of the test must be accomplished in the original choice of material of the test. Concerning the principles of the choice of material there is much confusion, particularly in subjects which are now in a period of development or transition. Should the test attempt to measure some typical course as represented by average textbooks or examinations used in the country or should material representing progressive experimental courses be selected? If standardized tests were used solely as measuring instruments, probably the former type would be more desirable, but the truth is that standardized tests also set up in the minds of teachers standards for the selection of material and instruction. For this reason they ought to represent in some degree forward-looking tendencies in curriculum development. Certain principles concerning the choice of materials may be safely given. Six distinct methods or criteria have been employed in the choice of material for standardized tests.

Criterion of Extrinsic Use

For spelling tests workers have been careful to choose words that are among the most common and most used in the English language. The Ayres spelling scale is really nothing more than a list of the thousand most common words in the English language, as determined by extensive word counts, whose difficulty for spelling is found by actual testing in schools. The word counts covered material in personal and business correspondence and standard literature and newspapers. Thus the selection is based entirely on the usefulness of the word in writing or reading

and not at all on its difficulty or peculiarity of spelling or any phonetic classification which has been the basis of so many spelling lists in the past. The same principle is extending to other fields. In algebra, for instance, due to studies of the use made of algebra in higher mathematics, in the applications of mathematics to science, and in technical reading, mathematics teachers have come to frown upon the complicated and intricate examples that occur frequently in older books. Fractions with polynomial denominators, complicated exponential and radical expressions are ruled out of the course. From these same people considerable criticism has come of examples found in standardized tests which no longer are sanctioned by various criteria of usefulness. This criterion, however, is difficult to employ unless the studies of frequency of use have already been made, for it is usually more difficult and tedious to make the studies than to construct a standardized test.

Criterion of Errors

Wilson [1] says concerning his Language Error Test:

> In 1909 the author, and his teachers at Connersville, Indiana, made a little study of the language errors of grade and high school pupils which showed that the ten most common errors accounted for over fifty per cent of all errors. Since then studies on a similar plan have been carried out in Boise, Kansas City, Cincinnati, and Iowa consolidated schools. With these studies as a basis, the author undertook, in 1916, to develop a language test that would distribute pupils on the basis of their ability to recognize and avoid the common language errors.

It is often difficult to know whether the criterion of selection of material for a test should be *use* or *errors*. Probably much depends on the emphasis in the course of instruction. If errors

[1] WILSON, G. M. — *Manual of Directions and Key* to Wilson Language Error Test, p. 1. Copyright 1923 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

are the basis of the language course, then the material of the test built to measure the outcomes of the language course may be selected on a basis of errors. If, on the other hand, the material of a course is new and unfamiliar to a class, if the emphasis is in learning new material and not in the correction of errors, bad habits, and the like, in familiar material, then both the course and the test to measure the course should be selected on the basis of frequency of use.

Composite of Textbooks

One of the most convenient and practical of the methods for selecting the material for a test is to use a composite of what is given in several textbooks in the subject. Concerning this criterion in the construction of the "Ruch-Popenoe General Science Test," Ruch[1] says:

> The selection of the items in the Ruch-Popenoe General Science Test has been based largely upon an analysis of the contents of existing textbooks in general science. Although there are more desirable criteria for the validation of test materials for most school subjects than the one here used, it is nevertheless true that, during the formative period of any new subject, teaching practice follows very closely the organization of the available textbooks. This very practical consideration is held to be sufficient justification for the construction of a test which embodies the best of the materials actually occurring in the books which must be used in general science classes. The original test consisted of three forms of fifty scientific terms each, which had been selected from a list of more than five hundred of such items obtained by analysis of twenty-three texts and manuals in general science.

Composite of Requirements in Course of Study

Very similar to the use of textbooks is the use of requirements in courses of study. Perhaps the latter are less definitely expressed for high school subjects than is the material in textbooks.

[1] RUCH, G. M., and POPENOE, H. F. — *Manual of Directions* to Ruch-Popenoe General Science Test, p. 1. Copyright 1923 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

Miss Strang [1] made use of courses of study in the construction of the " Gates-Strang Health Knowledge Test " in conjunction with the material in health textbooks.

Composite of Teachers' Examinations

This criterion was used effectively by Ruch-Cossman [2] in their biology test. The authors say regarding this:

Considerable study was devoted to the problem of securing valid content for inclusion in this test. In a high school subject which is as little open to standardization as is the case with biology, a great deal of preliminary investigation was necessary to determine exactly what is taught in good high school courses in biology. This lack of standardization of content has complicated the problem of validating material for test purposes. The selection and validation of the items in the Ruch-Cossman Biology Test have been largely based on the analysis and condensation of two thousand final examinations used in all parts of the United States over a period of several years.

The final examination represents a more or less thoughtful attempt on the part of the teacher to isolate those facts and principles which can with most reason be expected to be retained by the pupil for some appreciable length of time after the completion of the formal work of the classroom and laboratory. It is for this reason that the authors chose to attempt to survey good teaching practice in biology by statistical analysis of final examinations collected from all parts of the country. The limitations of the method are as apparent as are its advantages, and no claim is made that finality of content determinations can or should be based upon this method. The method of attack will be very briefly presented in summarized form.

(1) A letter was sent to the Superintendent of Public Instruction in every state of the Union having a central officer, asking that he list the names of those schools and teachers who were doing the highest grade work in high school biology. Lists were received from 23 states, with a total of 126 schools answering to our specifications.

[1] Strang, R. — *Subject Matter in Health Education: An Analysis and Evaluation of Some Courses of Study and Textbooks Dealing with Health, and Suggestions for Using Such an Analysis.*

[2] *Manual of Directions*, Ruch-Cossman Biology Test, pp. 1–2. Copyright 1924 by World Book Company, Publishers, Yonkers-on-Hudson, New York.

(2) Each of the 126 teachers was asked to send duplicate copies of all final examination questions used during the current year. Lists for previous years also were requested, and in some cases these were available for periods of five years or more prior to 1918. The total number of questions received ran slightly over two thousand.

(3) The two thousand questions were then analyzed for frequency and condensed and classified as 300 constantly recurring items. The group headings finally numbered eleven such phases of the total subject as: definitions of terms, identification of structures, life histories, diagrams, functions, economic and ecological facts, etc.

(4) The final list of 300 questions was then submitted to 100 of the leading teachers and authorities on the teaching of biology, for ratings as follows: (*a*) Rate " 1 " all questions which you consider to be entirely satisfactory and representative of the kind of knowledge which should be presented to high school students of biology, and which high school students can reasonably be expected to answer. (*b*) Rate " 2 " all questions which are partially satisfactory, but which you consider to be not highly desirable for any reason . . ., etc. (*c*) Rate " 3 " all questions which are entirely unsatisfactory.

Provision was also made for the writing in of any other questions not already listed. Sixty-eight teachers and 9 authorities (authors of texts, professors of the teaching of science, etc.) rated all the questions submitted. This was a remarkably high percentage of responses, and whatever selection did result from the failures to answer can safely be assumed to be a selection upward and hence could not have operated to decrease the reliability of the results.

Judgment of Experts

In all cases as a last resort we go back to the judgment of experts. Perhaps we should not say a last resort, because the judgment of experts has been the decisive factor in determining the content of textbooks and courses of study. It will be noticed that experts were called in to rate the desirability and suitability of topics and items for a high school course in biology. The material in the "Stanford Achievement Test," an excellent omnibus test in the elementary field, was largely selected from much more extensive material which had been selected through the combined judgments of T. L. Kelley, G. M. Ruch, and L.

M. Terman, its authors.[1] As an example, 2,500 elements of information in geography, elementary science, nature study, domestic arts, history, and literature had been assembled in connection with the study of gifted children in California. The original list of 2,500 items was reduced to 1,630 by the ratings of these judges and was then broken into three forms of 670 items each.

SELECTION OF TEST ITEMS

Makers of our best standardized tests have usually assembled at the start several times as many items or elements as it was their intention to keep permanently. It is generally recognized to-day that a standardized test should have at least two, preferably more, duplicate forms, so that the test may have its greatest usefulness in actual classroom practice. Often a teacher wishes to re-test her class soon after a first test is given. It is not desirable to repeat the identical test because certain of the items might be remembered and perhaps especially looked up in the interval. It may be affirmed here that there is less danger of this than is commonly supposed. However, it is usually possible and hence desirable to prepare and use tests identical in form but differing in content from the original test. The other reason for preparing many more test items at the start than it is intended to ultimately use in the test is in order to be able to discard items which seem to have little validity for the purposes of the test. Textbooks, courses of study, and teachers' examinations have been used to determine the most valid elements. Judgments of individuals have also been used to sift through the most valuable items. But if time and money permits, there is a still more objective method of selecting the most valid items from the total number assembled. Validity for test purposes does not always coincide with agreement with the objective of teaching. Oftentimes one item is better than another because it is better phrased or worded.

[1] *Manual of Directions*, Stanford Achievement Test, pp. 8–12.

In multiple-choice items sometimes one of the alternative choices contains such a unique association with the main part of the item that it is especially valuable in differentiating between the superior and inferior pupils. An example from the "National Intelligence Test" illustrates the point.

In each row draw a line under each of the two words that tell what the thing always has.

love (affection attachment display kisses sweetheart)

The last two items on the line are often underlined by adolescents.

Such obscure factors are practically impossible to apprehend except by trial. The problem is in reality a complex one and may strictly be resolved into two: (1) to find those items which independently contribute most toward predicting a criterion; (2) from these items select items at different levels of difficulty and arrange them in a test extending from as easy to as difficult as is desired.

Terman used this method in constructing his "Group Test of Mental Ability." In the manual of directions he says,[1]

The test as it stands is composed of questions and problems which were selected from a much larger number by correlating each separate item with a dependable measure of mental ability. The criterion used for this purpose was a composite which included grade, location, age, total score on a two-hour mental test, and ratings of the pupils by from two to five teachers on intelligence and quality of school work. The trial series was composed of thirteen tests with a total of 886 items. Try-out of these resulted in the elimination of three of the thirteen tests, and in the reduction of 610 items in the remaining tests to 370. All items which failed to differentiate pupils of known brightness from pupils of known dulness were eliminated. Since only the cream of original material was retained, it has been possible to reduce the length of the examination considerably below the limits which otherwise would have been necessary.

[1] P. 3.

Choose a Test because It Is Reliable

The second most important criterion of a test is its reliability. Indeed it is perhaps as hard to construct tests with the desired reliability as with high validity. Kelley states the principle, "If the testing time is of necessity brief, give prime consideration to reliability of the test, and if the testing time is long, give prime consideration to validity and but secondary attention to the reliability of the test."[1] To-day the maker of a good test reports the reliability of his instrument in the manual of directions accompanying the test or in some article in a technical journal which describes the test. One should suspect the reliability of a test whose author has not taken the trouble to determine its reliability.

Reliability refers to the correlation between the results of two forms of the same test. It is the accuracy of a test. Unreliability is best exhibited by giving an alternate form of a test and noting the change in score. A perfectly reliable test is one for which the score on a second trial remains exactly the same as on the first trial (barring practice effect). A perfectly unreliable test is one where the scores on two successive repetitions of a test are such as might occur by pure chance — such as one might draw out of a hat containing a number of possible scores. In estimating reliability we are not concerned with repeating the identical test, but in giving a test having the same construction, form, and subject matter as the original test but having different items. For example, the two forms of the "Powers Chemistry Test" are similar in all respects except that the elements are different. In using tests similar in nature but differing in content, one is really comparing the two halves of a test double the size of either form. One might easily make a test twice the length of the Powers Test by combining the two forms of the

[1] *Op. cit.*, p. 208.

test, item by item. Kelley calls two forms of a test comparable tests, and then defines comparable tests as tests having equal means and standard deviations, equal correlations with a criterion,

$$r_{01} = r_{02} = r_{03}, \text{ etc.}$$

and equal correlations with each other,

$$r_{12} = r_{13} = r_{23} = r_{1I}, \text{ etc.}$$

He also states that in testing reliability the chance differences between the two tests used to measure unreliability should not be correlated. If in a chemistry test a pupil does not know what NH_4Br stands for because of absence on the day of that lesson, we call the error a chance error or failure. If the identical test is repeated, the same failure is repeated and these chance errors or failures are correlated — the reliability is higher than the true reliability. But if comparable tests are given, these chance errors are practically uncorrelated.

That a student of tests understand thoroughly the meaning of reliability and its significance is highly important. Oftentimes the novice will consider scores as absolutely correct. When he inspects an individual's score on a test, he considers it a final and absolutely authoritative index of ability or achievement. Such a tester labels children as "an 86 I Q boy," or "his reading age is 145 months." One might permit such a point of view in the layman whose only experience with measurement is in connection with feet and pounds and where a pound is a pound and a foot a foot to the nearest integer or decimal place of measurement. Equally inexcusable in the novice at testing is the attitude so common that the results of testing are totally unreliable; that the boy who makes an I Q of 86 may be a genius in spite of the test result and that low achievement on a test is after all no true index of real ability. The truth lies somewhere between these two points of view — a test score is neither perfectly accurate nor is

it absolutely wrong. But just how accurate is a score on test X? The problem is largely one of suitable expression of this accuracy or reliability.[1]

FACTORS WHICH INFLUENCE RELIABILITY

Objectivity. — When the judgment of the scorer enters into the determination of the score, the test is called subjective. A subjective test must be unreliable, for it is obvious that the very same test paper may be scored differently by two different scorers or by the same scorer at two different times. Objectivity may be determined by inspecting the form of the test and noting whether judgment is necessary in the scoring. If one has to use considerable judgment in determining the correctness of an item, then the test is subjective. Ratings on an English composition are highly subjective. On the other hand, if only one correct answer is admissible for each item, and that one is to be clearly indicated in some way, then the test is highly objective. Objective tests are usually recognition tests as opposed to recall tests.

Examples

Recall Type: What is the most ductile metal?
Recognition Type: The most ductile of these metals is (1) iron (2) copper (3) platinum (4) zinc (5) lead

That our knowledge is used more frequently for recognition than for recall is probably true. In the long run we more frequently have to judge of the truth or falsity of statements that we hear or read than actually to provide the answers to questions. Whether or not one agrees as to the validity of objective type questions, the largest possible loss in validity is more than offset by the increase in reliability.

The matter of objectivity is one factor of unreliability which can

[1] See appendix B for a description of methods of determining reliability.

most easily be remedied. Of all the factors entering into unreliability, lack of objectivity is perhaps the most inexcusable, for it is usually possible, by exercising sufficient ingenuity, to turn a subjective test into an objective test with little or no loss in validity to the test. The test forms described in Chapter III are all examples of objective tests. The correct responses are carefully defined in a scoring key, so that all opportunity for individual judgment in scoring is removed. Usually in objective tests no part credit is given to any item — it is either right or wrong and the score on the whole examination is the sum of the separate credits on the elements. Unreliability arising from this factor may be entirely eliminated except as careless errors creep into the scoring.

Length of the test. — One of the most potent factors in determining reliability is the length of a test.

Two questions give more reliable information of a person's knowledge or ability than one. A test with fifty items gives a more reliable estimate of a person's ability than a test of twenty-five items. A limited sample of anything gives an untrustworthy, unreliable picture of the whole. Increase the sample and you coincidently increase the confidence that what you actually have is truly representative of the whole. If a boy fails one question in a five-question test, he is only able to show that he knows 80% of the subject — but in a 50-item test if he misses that same question it is only 2% of the total score and makes comparatively little difference in the result. The effect on reliability of increasing the number of questions is known mathematically — there is a definite and precise relation between the reliability of two similar tests of different lengths.

Evenness of scaling. — Not only should the items of a test be selected because of the fairness and adequacy of the sampling of the topic to be tested, but also because of their range of difficulty. By selecting the items in a certain way it is possible to

construct a test of twenty items so that it has no more reliability than a test of less than ten items.

Scale of difficulty

Scale of difficulty

Consider the test with items selected on the scale of difficulty in the first figure. If five of the items are so easy that no one fails on them and if five are so difficult that no one passes, then these items are practically useless for differentiating individuals, or for showing exactly what any one individual's ability is. If the other items are also closely bunched at two levels of difficulty, there will be a preponderance of scores of 5, 10, and 15 and thereabouts for those who cannot do items of the second level of difficulty, for those who can do the items on the second level but not on the third, and for those who can do those of the third level. Such a test might give a distribution with three modes. This would reduce the reliability to that of a test of between 3 and 10 items equally spaced in difficulty. There would be only three levels of scores as in the three-item test, but the scores at any level would be much more certain because it is the result of five trials at that level instead of one. If the items of a test are equally spaced in difficulty as in the second figure, there is no such loss in reliability due to coarseness of the measuring scale. The distinction is much the same as that of trying to measure length merely in feet or in feet and inches. The latter automatically give more accurate results. Unless one know definitely the difficulty of the items of a test there is a chance that the items will be bunched in difficulty, not in a regular way as pictured in the first diagram, but enough to materially affect

reliability. Wood found this true in his study of the C. E. E. B. examinations.

0	1	2	3	4	5	6	7	8	9	10
XX	XXXXX		X		X		X			

Wood says,

Let us suppose that the horizontal line represents the scale of ability in algebra, and that the arabic numbers and their distances one from another represent true, equal, and uniform increments of ability. The distances marked off by the "×" signs are rough projections of the difficulties (measured in two ways) of the ten elements in the Mathematics A June 1921 examination, each of which ten elements was given identically the same incremental credit value of 10% of the total possible score. The possible mischief from such a scale is enormous. For example, let us suppose that those who jump the hurdle nine feet high should pass, and those who do not succeed in getting over it should fail. Now it is obvious that the ninth "×" in the figure is much more than one foot or one unit higher than the hurdle which we call eight feet high, and therefore the real nine-foot hurdle lies somewhere between the eighth and ninth stars. Now this increment is so great in comparison with the other increments that it might well be that 50% of our total candidates really can jump an altitude between the eighth and ninth stars; that is, many candidates can jump nine feet, but cannot jump the ninth hurdle. This results in a large undistributed group at the eighth star which really belongs all way between the eighth and ninth stars.

Condition of the pupils taking the test. — McCall[1] divides the causes of test unreliability into factors inherent in the test itself and factors inherent in the pupils taking the test. The three factors already discussed have been mechanical factors in the test itself. The other factors have to do with the conditions under which the test is taken and the variations in the pupils taking the test. With regard to the latter, there is much misapprehension. The instability of the individual taking the test is commonly regarded as the outstanding factor in causing test unreliability. Most people feel that strain or nervousness or a physical condition not quite the best works havoc with their performance on an

[1] McCall, W. A. — *How to Measure in Education*, pp. 307, 308.

examination. Not only does popular impression attribute much of the cause of unreliability to this nervousness, etc., but even experts in the field of testing make similar statements. Courtis says, "The best examination is not that represented by the score of a single performance in a single day. Human effort is variable and human skill too easily upset to make it fair to have promotion based upon chance scores." Kelley[1] says, "At a single sitting one's buoyancy and vigor in attacking the odd problem is probably much the same as that in attacking the even problems. A feeling of discouragement because of failure on a problem probably carries over to the next problem, thus increasing the chances of failure on it. One's general tone, up or down, as a result of a good sleep, a poor breakfast, etc., is the same on the odd as on the even problems." Kelley indicates by this that he believes such phenomena have a *certain* amount of influence on test reliability. McCall[2] says, "An automobile horn, the lonesome howl of Jack's dog, the bleating of Mary's lamb, a sudden thought of the swimming hole, growing discomfort of a strained posture, these and a thousand other large and small internal and external influences register themselves in the pupil's scores." In a study of reliability I[3] give evidence to show that the general conditions of an individual exert practically no influence under ordinary school conditions on the results of testing. On the contrary, not one of 232 school children showed any appreciable unreliability in tests due to this personal factor. Two scales of the "National Intelligence Test," both Form I, were administered on a given day and a week later the tests were repeated, using Form II of both scales. The hypothesis was that if the general condition of an individual is a factor in test unreliability, then a child who was feeling ill or nervous on

[1] Kelley, T. L. — "Note on the Reliability of a Test: A Reply to Dr. Crum's Criticism," *Journal of Educational Psychology*, 15: 193–204, April, 1924.

[2] *Op. cit.*, p. 308.

[3] Symonds, P. M. — "A Study of Extreme Cases of Unreliability," *Journal of Educational Psychology*, 15: 99–106, Feb., 1924.

the second day, as compared with the first day, would make a much lower score on several of the sub-tests in the scales on the second day than on the first day. The following tables tell the story.

TABLE 80

NUMBER OF INDIVIDUALS MAKING EXTREME PLUS OR MINUS DEVIATIONS OR REPETITIONS OF THE NATIONAL INTELLIGENCE TESTS

Number of Tests with High Plus Deviations on Second Trial N. I. T. A.	Number of Individuals
0	174
1	52
2	5
3	1

Number of Tests with High Minus Deviations on Second Trial N. I. T. A.	Number of Individuals
0	178
1	45
2	8
3	1

Number of Tests with High Plus Deviations on Second Trial N. I. T. B.	Number of Individuals
0	191
1	48
2	3
3	0

Number of Tests with High Minus Deviations on Second Trial N. I. T. B.	Number of Individuals
0	195
1	43
2	3
3	1

8 individuals had both a high and low deviation on A.
9 individuals had both a high and low deviation on B.

This means that only one pupil made 3 increases in score on his second trial that could be called large increases. (There are five sub-tests all together in form A.) Only five children made increases on two sub-tests that could be called large. Eight pupils made a large increase on one sub-test and a large decrease on another sub-test. One is as apt to make a large increase as a large decrease on a repetition of a test. In this study there was no evidence that nervousness, state of health, or enthusiasm had any appreciable effect on scores. This experiment needs to be repeated because the results so contradict both popular and expert opinion. Woodyard [1] finds that test reliability is as high with the second test given immediately after the first as when a year intervenes. Evidently the cause for test unreliability must be sought elsewhere than in the general condition of the individual. This does not mean that one should discount entirely such factors. A boy who breaks his pencil and is unable to finish his test manifestly does not do himself justice. So a person who is forced to write with his left hand because of injury to his other hand, or who is so sick as to be unable to attend properly, does not do himself justice in an examination. But the contention is that under ordinary school conditions such occurrences are rare enough not to appear in our statistics — and the statistics show much unreliability. It is remarkable how, under the stress of taking a test, an individual can pull himself together to do his best. This seems to be the general rule rather than the opposite.

Familiarity of pupils with the technique of taking tests. — The investigation above mentioned, "A Study of Extreme Cases of Unreliability," showed that much of test unreliability is due to lack of training in the technique of taking a test. For instance, there were great variations in the speed with which the tests were taken. In tasks which were unfamiliar the pupil would often seem

[1] Woodyard, E. — *The Effect of Time upon Variability*, Teachers College Contributions to Education No. 216, 1926.

to be struggling with what to do the first time, but once he knew what was expected he would go much faster the second time with a clear understanding of the task. Sometimes just the opposite would take place. On the first trial the pupil would commence underlining rapidly with no clear conception of the problem. Later he would catch on to what was expected and the new job of actually doing the problems would slow him down.

Differences in accuracy is another cause of unreliability.

In every case a high plus deviation has been accompanied by increase in accuracy, and a high minus deviation has been accompanied by decrease in accuracy. The same factor we noticed in considering speed is evident here — "catching on" to the problem leads to greater accuracy. In the case of high minus deviations accuracy is also a factor. My notes read as follows: "hurried — careless," "harder — tendency to let go of set," "faster — careless toward end," "less check on self."

In conclusion I would like to suggest the possibility of lowering test unreliability by means of systematic training. Pupils should be taught to take tests. In the first place pupils need to be taught to go through the motions that modern testing demands. They need to be taught to put down answers on the dotted line, that they may figure on the sides of the paper if need be, what it means to underline, or cross out, to follow directions if they are told to underline only one word or two words, to write letters where they should be written, etc. Specific training should be given.

Then again, definite training needs to be given in the taking of tests themselves. A pupil needs to learn that getting an example right is the only thing that scores. He needs to be taught that he is not expected to finish the test but to do as much as he can do *well*. He needs to be taught that getting a certain percentage on the test is not what counts, but what he does in comparison with others of the class. Perhaps the best training in this is to show the pupils their scores and let them see where they stand with respect to others in the same class.

RELIABILITY OF NEW-TYPE EXAMINATIONS

The reliability of traditional examinations and new-type examinations is a matter of much interest. One of the talking points for the new-type examinations is their superior reliability over

their predecessors. Wood finds that the College Entrance Examination Board Mathematics A (Algebra) examination in 1921 had a reliability of .76. The Mathematics C examination (geometry) is .61. Monroe [1] found 66 reliability coefficients of traditional examinations given in both elementary and secondary school subjects in single classes. The average reliability coefficient was .65 and this may be taken as a typical figure for the reliability of traditional examinations.

Wood,[2] Toops,[3] and Ruch [4] have studied the reliability of new-type tests with results as follows:

TABLE 81

Reliability Coefficients Reported by Wood, Toops, and Ruch for New-Type Tests

Reported by	Type Test	No. Statements or Questions	Time	Reliability
Wood	True False	139	70 minutes	.89
	Completion paragraphs	14	35 minutes	.89
	Total test		2 hours	.95

	Type of Test	Reliability, 100 Items	Reliability, Equal Time (Time to Complete 100 Recall Items)
Toops	Recall	.764	.764
	5 Response-multiple choice	.715	.755
	True False	.673	.798
Ruch	Recall	.896	.896
	5 Response-multiple choice	.886	.901
	True False	.714	.820

[1] Monroe, W. S., and Sanders, L. B. — "The Present Status of Written Examinations and Suggestions for Their Improvement," *University of Illinois Bulletin, Vol. 21, no. 13*, p. 34, 1923.

[2] *Op. cit.* p. 193.

[3] Toops, H. A. — *Trade Tests in Education*, Contributions to Education, no. 115, Bureau of Publications, Teachers College, Columbia University, New York, 1921.

[4] Ruch, G. M., and Stoddard, G. D. — "Comparative Reliabilities of Five Types

Ruch's coefficients are higher than Toops', probably because Toops' were found from testing in college classes, while Ruch's were from testing in high school classes.

The evidence is decisive that new-type examinations are more reliable than traditional examinations. In fact, the evidence is more decisive than an off-hand inspection of the above tables would indicate. One usually wishes to compare examinations for reliability under equal working times. Monroe does not state the average length in time of the 66 sets of examinations used by him in getting his reliability coefficients. The median number of questions in an examination was 7, and one may assume that the examination time was at least an hour (the C. E. E. B. examinations allow 3 hours). Ruch's correlation of .896 for 100 recall questions was based on 18.7 minutes of testing time. Thus it is a matter of comparing an average reliability of .65 for the traditional examination obtained in 60 minutes more or less of testing time with a reliability of .90 for a new-type test obtained in 19 minutes of testing time. The superiority of the new-type test, as far as reliability goes, is plainly evident.

WHAT SHOULD ONE EXPECT IN THE WAY OF RELIABILITY?

One answer is: expect the highest reliability you can get in the time allowed, — that is, set the time you can afford for testing and then use that type of test which will give you the highest reliability. Much depends on the number of items that can be asked in the time allowed and also the validity of these items. Ruch finds that he can ask about 188 5-response multiple-choice exercises in 30 minutes, yielding a reliability of about .935. Brinkley[1] finds he can give only 56 multiple-choice exercises in 30 minutes.

of Objective Examinations," *Journal of Educational Psychology*, 16: 89–103, Feb., 1925.

[1] BRINKLEY, S. G. — *Values of New Type Examinations in the High School*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 161, 1924.

Using Ruch's data, Brinkley's 30-minute multiple-choice test would yield a reliability of .814. The answer to the question: What reliability should one expect? must remain a qualified one. The factors which enter into unreliability are so many and obscure that tests may fail to show high reliability for unknown reasons. With thirty minutes of testing where 50 or more objective questions are asked one should expect to get over .80 for a reliability coefficient. With increased testing time involving a larger number of questions the reliability rises to any reliability desired. One gets in test reliability what one is willing to spend in testing time. When one considers how poor a prediction is obtained from a correlation of .85, a test occupying 45 minutes seems hardly adequate in such subjects as general science, biology, or chemistry. Teachers should come to realize that in order to make testing worthwhile, the test results should be reliable. This implies continuing the test over two or more school periods. One should aim to get a reliability of .90 in all cases where the results of the testing are to be used seriously in the administrative or guidance functions of the school.

What do we know about the reliability of standardized tests? We know that on such a test as the Powers Chemistry Test or the Ruch-Cossman Biology Test a score of 15 may not be the true measure of a pupil's ability in terms of the units of the test. But we know that it is not far from 15. There is an even chance that if a pupil earns a score on the test, his real ability will be represented by scores between 12 and 18. There is practically no possibility that a boy who makes a low score deserves a high score. Scores on standard tests are not absolutely accurate, but they are accurate enough so that we may place considerable confidence in test results.

The paucity of determinations of reliability is pathetic. The tests for which reliability coefficients or probable errors of a score can be found are a mere handful. No better evidence of the youth

of the testing movement in high schools exists than this fact that workers can assemble a test and publish it without demonstrating that the test is better than someone's else. The time will surely come when teachers will be trained to demand tests of a specified reliability.

Choose a Test because It Is Scaled

A scaled test is one in which the items of the test progress according to some quality such as difficulty, æsthetic quality, legibility, etc. A test scaled on the basis of difficulty is one in which each item is more difficult than the preceding one. Most standardized educational tests are scaled on a basis of difficulty. In order to make a test a true scale in the same sense that scale is used in other branches of science, two other features are essential. One, the items of the test should proceed with equal increments of difficulty. That is, item 3 of the test should be as much more difficult than item 2 as item 2 was more difficult than item 1, and so on. The other qualification is that zero on the scale should represent "just not any" of the thing which is being measured by the test. Difficulty, in the sense in which it is used above, may be determined in two ways: (1) by the combined judgment of a number of competent people, (2) by the decreasing percentage of a group taking the test who can answer successive items. Thorndike[1] has shown that there is a high correlation between these two procedures in determining difficulty. The second method is to be preferred when it is possible.

There has been much misapprehension concerning the scaling of tests. In general, the scaling of a test has been much overemphasized in past discussions. It is surely important that handwriting, composition, sewing, and drawing scales be accurately scaled. In the past, however, much effort has been spent on the

[1] Thorndike, E. L., Bregman, E. O., and Cobb, M. V. — "The Selections of Tasks of Equal Difficulty by a Consensus of Opinion," *Journal of Educational Research*, 9: 133–139, Feb., 1924.

scaling of a test which might more profitably have been spent in making the test thoroughly valid and reliable. A decade ago test makers assigned each item of a test a scale "value" and scored the test by securing these scale values. This procedure was costly in time and labor and easily led to errors. Later it was discovered by several workers independently that the number of correct items on a test and the sum of these scale values correlated well over .90, in fact, so high that it was admissible to use the number of correct items as a measure of pupil achievement in place of the more cumbersome "sum of scale values." To-day "number of correct items" is nearly universal as a scoring formula. In fact, providing certain general cautions are observed, it is permissible to merely arrange the items of which your test is to be composed in order of difficulty without paying much further attention to scaling. One caution to be observed is that there should be some items on the test so easy that no pupil will get a zero score. A zero score is called an undistributed score. By this is meant that the scores fail to show individual differences. Two individuals with zero scores may really differ in the ability in question, yet this difference will not appear in the score. A test with zero scores is too hard. Likewise there should be some elements in the test of such a difficulty that no pupil in the class will get a perfect score. Perfect scores on a test are also undistributed scores and equally as pernicious as zero scores. Perfect scores fail to distinguish between individuals of high ability. The introduction of more hard elements into a test which yields perfect scores would undoubtedly produce a spreading out of the perfect scores. There should be no undistributed scores at any point of the scale. Such undistributed scores sometimes occur in the middle of a test when items are bunched in difficulty as described in page 292. A final qualification is that the test should be of a general difficulty such that the mean score of a class is equivalent to about 50% of the total possible score on the test. With these qualifications it

makes little difference whether or not the items are accurately scaled for general school usage.

Most standardized tests for use in high school have such a generous time limit that speed is not a factor in the score.

Of course it should be understood that a test which is not thoroughly scaled is useless for experimental purposes (unless highly refined statistical procedures are used). In experimental education one wishes to be able to say that a gain of 5 points from a score of 15 to 20 on a scale is equivalent to a gain of 5 points from a score of 25 to 30, which is only true when items proceed by equal increments of difficulty. Likewise one would like to be able to say that a 20% gain of a score from 5 to 6 is equivalent to a 20% gain of a score from 25 to 30, which is only true when a zero score is a true zero. In the use of intelligence tests workers have been much handicapped by the fact that there has been no universal, truly scientific scale of intelligence. The mental age scale has been most unsatisfactory in that it is not known whether it is a linear scale in its lower ranges and it has broken down completely as a scale in the upper levels of intelligence. Thorndike's [1] recent work in deriving an absolute scale of intelligence meeting all technical requirements of a scale is to be reckoned as a truly great achievement.

Other Things Being Equal, Choose a Test That Has Norms

A standardized test without norms is practically useless. Enough has already been said in Chapter XIII to show the importance of norms. Norms or standards are essential for interpreting the results of a test. A class average should be compared with the norm or average performance of other groups. A pupil's score can only be satisfactorily interpreted in the light of percentile norms or some similar set of indices which will permit comparisons of a pupil's performance with performances of a

[1] Thorndike, E. L. — *The Measurement of Intelligence*, 1926.

large number of pupils. Norms should accompany a standardized test. Most publishers include the norms in the manual of directions accompanying the test or on the tabulation entry sheet which they sometimes provide. Select tests for which there are satisfactory norms derived from a large number of cases. This is evidence that the test has been widely used.

Other Things Being Equal, Choose a Test That Is Inexpensive

This criterion and the two following do not represent actual test superiority, but are merely criteria that answer administrative exigencies. Probably no teacher or school administrator needs to be told to select an inexpensive test. There is little or no correlation between the price of a test and its value as a test. Some years ago workers at Indiana University endeavored to break the "test trust" by distributing tests at a very low cost. Unfortunately the tests themselves were not prime tests and the movement seems to have died out. The tests were printed on low-grade paper and had the appearance of being cheap tests in all senses of the word. To-day publishing houses issue series of standardized tests, selected and edited by expert psychologists, print them in an attractive way on good quality paper, and charge prices commensurate with their quality. Previous criteria are more important than this one of cost and one should always choose the best test from the point of view of validity and reliability, even if high priced.

The lowest-priced test is of the type of the "Ayres Spelling Scale." The "Ayres Spelling Scale" is not really a test at all, but merely a list of one thousand words whose difficulty has been carefully determined. Words of like difficulty have been put in columns and the norm of each column for all grades of the elementary school is given. With the Ayres Scale one may

easily select words to make his own test. What we need in the field of testing is lists of valid elements of known difficulty in each subject, from which a teacher could readily construct her own test, having the optimum validity for her class and the desired reliability. With such lists available a city system could construct and print its own tests, thereby freeing itself from dependence on publishing houses and reducing costs to a minimum. Test making is still in its infancy, and when workers are through scratching the surface and producing merely usable standardized tests, they will proceed to dig the ground deeper and produce exhaustive lists of elements together with their difficulty and validity at various stages of learning.

Other Things Being Equal, Chooose a Test That Is Easy to Score

This criterion is also a sop to administrative exigency. A test should not be primarily selected because it is easy to score, but at the same time a good test ought to be as easy to score as possible. Publishers usually distribute scoring keys with standardized tests. The key should be printed as near to the margin of the paper as possible so that it can be brought close to the pupil's answers. It should be spaced exactly as the material is spaced. Otis has adopted the principle of placing parentheses at the right-hand side of the test sheet. The pupil writes in each parenthesis the number which indicates his answer to the question. This makes it possible to bring the answers and the key into close conjunction. Multiple-choice questions ought always to be of the type

The Forsyte Saga was written by (1) Howells (2) Wells (3) Churchill (4) Galsworthy (5) Hutchinson (4)

rather than the type

The Forsyte Saga was written by Howells, Wells, Churchill, <u>Galsworthy,</u> Hutchinson.

The former causes much less eye strain in scoring.

Sometimes a stencil is provided as a scoring key, as in the "Briggs Form Test" and the "Rogers Test of Mathematical Ability." A stencil sheet is a sheet with holes perforated indicating the place where the answer is to be expected, such as a word crossed out or underlined. Usually a stencil is saving of labor, but undoubtedly many errors escape a scorer using a stencil. For instance, if a pupil underlines two words in a multiple choice test (one of which is correct), the underlining of the second word would not be seen with a stencil. Usually items in which two or more words are underlined are given no credit.

Other Things Being Equal, Choose a Test Having Alternative Forms

The best standardized tests are made with two or more duplicate forms, and test users should insist on using tests that have alternative forms for several reasons. If it is desirable to repeat a test during the course of the year's work, there will be no danger if previous test items have been looked up in the meantime and remembered. Another reason is that oftentimes a teacher wishes to measure his class with greater reliability than a single test affords and alternative forms provide an easy way of immediately extending the length of a test. The author of a test needs at least two forms to determine the reliability.

When standardized tests were first marketed, authors put the best and most valid material in one test. Now that enough material must be assembled to cover the requirements of two or more comparable tests, the best material is sorted out equally to all the different forms. For this reason when there are several alternative forms, no one form can contain the most significant or most valid material. One has to dip down in the reservoir for less important material. In intelligence, vocabulary, and reading this does not result in lowering the validity of the test to any appreciable degree — there is an inexhaustible stock of material

to draw upon. But Wood found in drawing up material for college entrance tests in ancient history that in order to assemble enough material for several tests, he would have to resort to asking questions of little importance and significance, questions which, from the pupil's point of view, are unfair, and from the testers, invalid. It is impossible to make many alternative tests in a subject of limited content. There must always be a mean between the desire to obtain a valid test and the desire to provide for increased reliability by supplying several alternative forms.

Duplicate forms should be equal in difficulty at all points of the scale. That is, a score of 10 on form *A* should be the equivalent of 10 on form *B* and a score of 50 on form *A* should be equivalent to a score of 50 on form *B*. This equating may be accomplished in two ways. If the items are scaled, then each item has a scale value. From the large reservoir of items assembled successive items which have the same scale value may be selected for inclusion in the test. A simpler method would be to select items answered correctly by the same percentage of a group as being of equal difficulty and to build up comparable forms directly from these percentages.

Other Things Being Equal, Select a Test Having Standardized Directions

This criterion of a good test is probably less important to-day, at least in the high school, than when McCall[1] wrote in 1922, "Long instructions so tax the memories of pupils that absolute essentials are frequently forgotten. To forget a single one of these essentials may markedly alter the child's score." Directions are particularly needed in speed tests, but in our ordinary content examinations in the high school, directions may be reduced to a minimum with no derogatory effect upon the score. Also it is probably true that variations from the exact printed directions

[1] *Op. cit.*, pp. 235 ff.

have less effect than was once commonly supposed. The directions for giving a standardized test in high school should cover the following points:

1. Statement as to what it is a test of.
2. Statement as to how many parts the test has if the class is allowed to go straight through.
3. Statement as to how many questions there are in a given part of the test if the class is stopped at the end of each part.
4. Warning to turn the page if the test continues on the next page.
5. Warning not to dwell too long on any one item or question.
6. A sample exercise which the examiner usually reads aloud.
7. Statement as to the mechanical form of answering — whether underlining, crossing out, writing a number in a parenthesis, etc.
8. Directions as to what to do in case a pupil finishes the test before time is called.

In conclusion one may summarize by stating that the two fundamental criteria for choosing a test are validity and reliability, — all others are subordinate. Objectivity in scoring is one important factor in reliability. A scaled test is to be preferred, although for actual school practice merely placing the items in an order of difficulty suffices in a comprehensive objective test. For experimental work, however, a test should be accurately scaled. In selecting a test one must be careful to see that it goes to low and to high enough limits. To-day the best standardized tests have accompanying them standardized directions and norms and are provided in two or more alternative forms. Administrative exigencies also add the two criteria of low cost and ease of scoring.

REFERENCES

BRINKLEY, S. G. — *Op. cit.*, Ch. 3.

CHASSELL, C. F., and CHASSELL, L. M. "A Survey of the Three First Grades of the Horace Mann School by Means of Psychological Tests and Teachers' Estimates, and a Statistical Evaluation of the Methods Employed," *Journal of Educational Psychology*, 12 : 243–252, May, 1921.

COURTIS, S. A. — "Report of the Standardization Committee National Association of Directors of Educational Research," *Journal of Educational Research*, 4 : 78–80, June, 1921.

CRUM, W. L. — "Note on the Reliability of a Test, with Special Reference to the Examinations Set by the College Entrance Examination Board," *American Mathematical Monthly*, 30 : 296–301, Sept.–Oct., 1923.

FRANZEN, R., and KNIGHT, F. B. — "Criteria to Employ in Choice of Tests," *Journal of Educational Psychology*, 12 : 408–412, Oct., 1921.

GATES, A. I., and STRANG, R. — "A Test in Health Knowledge," *Teachers College Record*, 26 : 867–880, June, 1925.

HOLZINGER, K. J., and CLAYTON, B. — "Further Experiments in the Application of Spearman's Prophecy Formula," *Journal of Educational Psychology*, 16 : 289–299, May, 1925.

JORDAN, A. M. — "The Validation of Intelligence Tests," *Journal of Educational Psychology*, 14 : 348–365, 414–428, Sept., Oct., 1923.

KELLEY, T. L. — "The Applicability of the Spearman-Brown Formula for the Measurement of Reliability," *Journal of Educational Psychology*, 16 : 300–303, May, 1925.

— "Note on the Reliability of a Test: A Reply to Dr. Crum's Criticism," *Journal of Educational Psychology*, 15 : 193–204, April, 1924.

— *Statistical Methods*, ch. ix, xii. *Ref.* Ch. 12.

KELLEY, T. L., RUCH, G. M., and TERMAN, L. M. — *Manual of Directions* — Stanford Achievement Test. World Book Company, Yonkers, New York; 63 pp., 1923.

McCALL, W. A. — *How to Measure in Education*, Part Two. *Ref.* Ch. 3.

MONROE, W. S. — *Introduction to the Theory of Educational Measurements*, ch. ix. *Ref.* Ch. 3.

MONROE, W. S., and SANDERS, L. B., "The Present Status of Written Examinations and Suggestions for their Improvement." *Ref.* Ch. 2.

OTIS, A. S. — *Statistical Method in Educational Measurements. Ref.* Ch. 12.

ROBACK, A. A. — "Subjective Tests versus Objective Tests," *Journal of Educational Psychology*, 12 : 439–444, Nov., 1921.

RUCH, G. M. — *Improvement of the Written Examination*, ch. ii. *Ref.* Ch. 2.
— "A Range of Information Test in General Science." *Ref.* Ch. 7.
RUCH, G. M., and COSSMAN, L. H. — "Standardized Content in High School Botany." *Ref.* Ch. 7.
RUCH, G. M., and STODDARD, G. D. — "Comparative Reliabilities of Five Types of Objective Examinations." *Ref.* Ch. 3.
RUGG, H. O. — "Standardized Tests and the Improvement of Teaching in First Year Algebra." *Ref.* Ch. 6.
— *Statistical Methods Applied to Education*, ch. viii. *Ref.* Ch. 12.
SPEARMAN, C., and KRUEGER, F. — "Die Korrelation zwischen verschiedenen geistigen Leistungsfähigkeiten," *Zeitschrift fur Psychologie*, 44 : 50–114, 1906.
SYMONDS, P. M. — "A Study of Extreme Cases of Unreliability," *Journal of Educational Psychology*, 15 : 99–106, Feb., 1924.
TERMAN, L. M. — *Manual of Directions* — Terman Group Test of Mental Ability. World Book Company, Yonkers, New York; 12 pp., 1924.
TRABUE, M. R. — *Measuring Results in Education. Ref.* Ch. 4.
WILSON, G. M. — "Language Error Tests." *Ref.* Ch. 5.
WOOD, B. D. — *Measurement in Higher Education*, chs. v-viii. *Ref.* Ch. 2.
— *The Reliability and Difficulty of the College Entrance Examination Board Examinations in Algebra and in Geometry.* The College Entrance Examination Board, New York; 33 pp., 1921.

CHAPTER XV

DERIVED SCORES

Up to this point the discussion has been about scores on tests. These scores were, in most cases, the number of items which were done correctly. On an English test of 30 items, for instance, if 24 items are correctly answered, the score is 24. These scores were interpreted by comparing them with norms. This chapter discusses some derived scores or units which have been suggested by various workers to make more meaningful the results obtained on standardized tests. At the outset let it be understood that these derived scores do not contain information which is not in the original scores, but they help us to interpret our original scores.

MENTAL AGE (M A)

The first set of derived scores is the ages, the best known of which is the mental age. Mental age is the score one obtains from the Binet test. The Binet test contains 6 tests on each age level. If a child is able to pass all the tests on the 8-year level, he is said to have a mental age of 8 years. The score on the Binet test is given directly in terms of mental age. The score on a group test, however, is the number of elements that are answered correctly. It is often desirable to turn these scores into corresponding units of mental age. In order to do this one must build a table by which it is possible to transmute scores into units of mental age. To construct such a table, one must find the average score made by a large number of 8-year-old children coming from different parts of the country, — cities, towns, and rural districts, — and also the average score made by children of other ages. It is

possible to read from the score the corresponding age. Mental age may therefore be defined as the age of the average child having the same score as that made by the child under consideration. For instance, if a 12-year-old child makes a score equal to the average score made by all 10-year-old children, we say that he has a mental age of 10 years. If he makes a score equal to the average score made by all 14-year-old children, we say that he has a mental age of 14. Mental age is determined solely by what a child does on a mental test. For the "Otis Higher Self-Administering Test of Intelligence" a table of age norms is given (p. 66).

To use this table one first finds the score and then reads the corresponding age; that is, a child receiving a score of 23 has a mental age of 12 years. Some writers call a mental age found in this way a group mental age (G M A) to distinguish it from the mental age as found on the Binet scale.

It should be noted in passing that the mental age obtained from one intelligence test may not be the same as the mental age obtained from another; that is, mental age obtained from the Otis Intelligence test would not necessarily be the same as that obtained from the Binet test. This is because the tests are not constructed of the same material, and because they are not completely reliable.

The following graph shows the increase of mental age with the growth of the child. Note that the growth is very regular until about the age of 14 or 15 when it begins to taper off. The average person will make no higher score on a mental test at any period of his life than he does when he is 17 or 18 years old. That we find boys and girls in high school and in college making higher scores than average boys and girls in the 8th grade who are 14 is not entirely because the former have continued to increase in their ability to do the things required on the test, but because their less able classmates have dropped out of school, leaving in school only those who are able to make the higher scores. During the period

of growth mental age increases regularly and gradually with no periods of marked rapid growth followed by periods of little growth as we were told by psychologists of an older generation.[1]

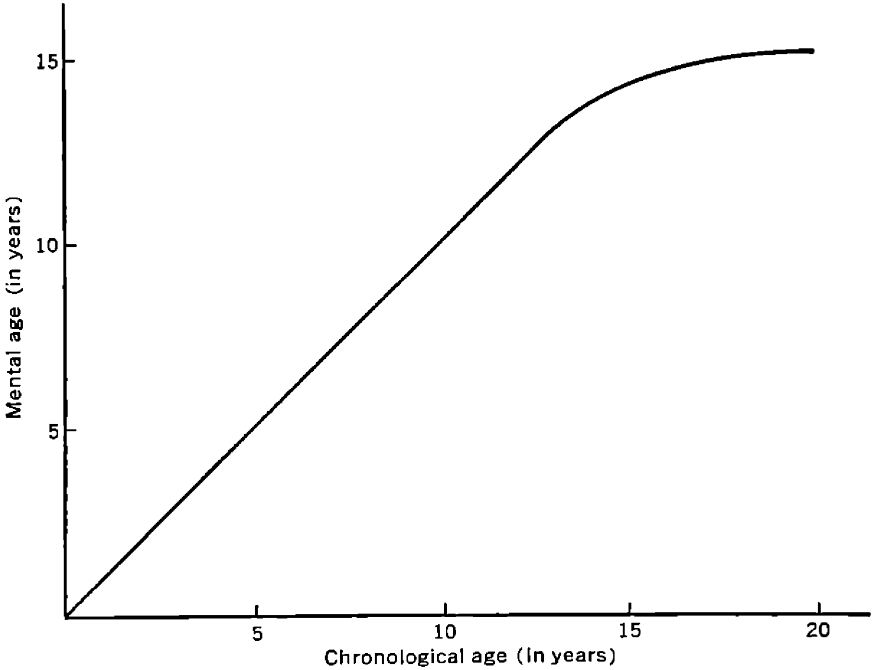


FIGURE 8. — Curve of mental growth

[1] The precise nature of the curve of mental growth after the age of 12 is a matter of conjecture and dispute. For a while after the war it was thought that mental growth ceased early — that is, at the age of 14 or 15. The work of Brooks (*Changes in Mental Traits with Age; Determined by Individual Retests*) confirmed the data of some previous workers that mental growth continued long after the age of 14, probably to 18 or 19. Thorndike ("On the Improvement of Intelligence Scores from Fourteen to Eighteen," *Journal of Educational Psychology* 14 : 513–516, Dec. 1923), more recently has demonstrated that there is continued growth in mental ability throughout the high school period of such a magnitude as to be "equivalent to at least 10 months of mental age around 14." The general impression seems to be that represented by the curve pictured on this page, regular growth up to the age of 13 or 14, followed by regular decreasing increments of yearly growth up to the age of 18 or 19. This growth, as nearly as our present instruments can measure, is regular. There may be sex differences at the

The symbol for mental age is M A and one should interpret M A to mean mental age wherever he finds it in the rest of this book. It should be evident to any one who carefully considers the method by which mental age is determined and the laws of mental growth that mental age is a very inadequate unit by which to measure mental ability or mental level. For, consider the bright child

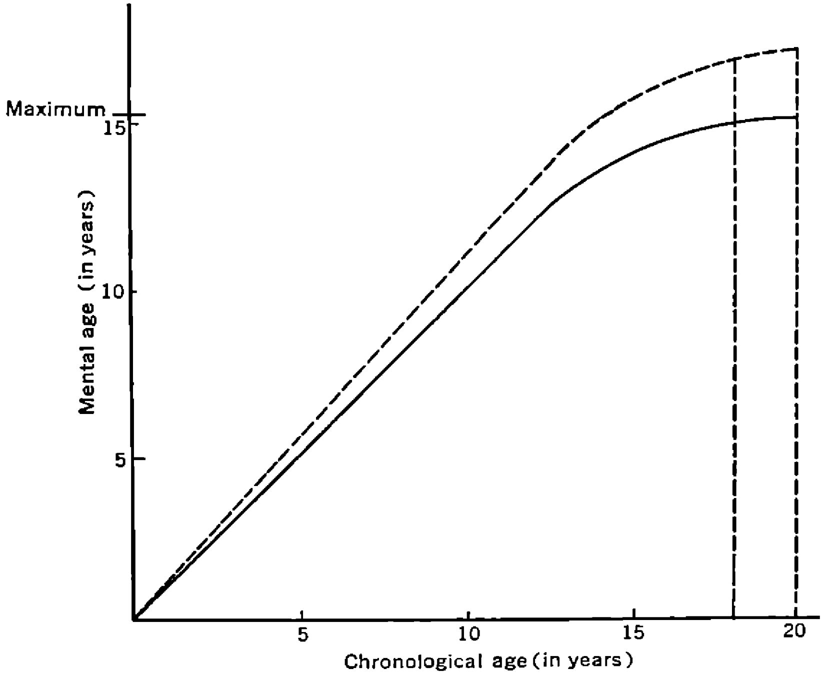


FIGURE 9. — Curve of mental growth of average and superior children

who grows more rapidly than the average child and reaches a mental level which surpasses the maximum ever reached by the average child. At the age of eighteen there is no score made by the average child of any age which equals his score. Suppose that the score made by our bright child on the Otis test above re-

age at which mental age begins to slow down, but our tests are not as yet refined enough to exhibit such differences.

ferred to is 50. In the table on page 66 there is no entry for a score of 50, — no entry for a score over 42, for the reason that the average child of any age never makes a score above 42. The score of 50 has no age equivalent. It is as though one wanted to weigh a package for the mail of approximately 20 pounds on letter scales with a maximum capacity of 15 pounds. Neither is the mental age scale of use for all children brighter than average and over 16 years of age. The Stanford Revision of the Binet Scale has surmounted this difficulty by arbitrarily adding tasks or exercises on the higher levels. Six tasks are designated as "average adult" tasks and six tasks as "superior adult." Passing the average adult tests will give one a mental age of 16 years, 6 months, while passing the superior adult tests (and all others below) will give one a mental age of 19 years, 6 months. Although Terman [1] states that he determined average adult level by testing 30 business men, 150 "migrating" unemployed men, 150 adolescent delinquents, and 50 high school students, there is scanty evidence that the tasks on the "average" and "superior" adult levels continue the mental-age scale in any kind of linear or other defined manner. The differences between "mental ages" derived from averages and the mental ages of the Binet test are well illustrated in Table 82.

About the age of 15 the mental age equivalent in score (57) begins to pull away from the age norms (53) and the divergence increases with each successive age.

Because age norms are totally inadequate for use with all superior individuals over the age of 14 and 15, mental age equivalents are sometimes used with the clear understanding that they are artificial extensions of the mental-age scale and are not based on averages.

Mental age as a unit may be safely ignored in the high school. There is no advantage to be gained by turning the scores into

[1] TERMAN, L. M. — *The Measurement of Intelligence*, p. 54, 1916.

TABLE 82

OTIS SELF-ADMINISTERING TEST OF MENTAL ABILITY: NORMS

INTERMEDIATE EXAMINATION

Age Norms		Binet Mental Age Equivalents	
Age	Score	Age	Score
8	7	8	9
9	15	9	16
10	23	10	23
11	31	11	31
12	38	12	38
13	44	13	45.5
14	49	14	51
15	53	15	57
16	56	16	61.5
17	58	17	66.3
18	59	18	71

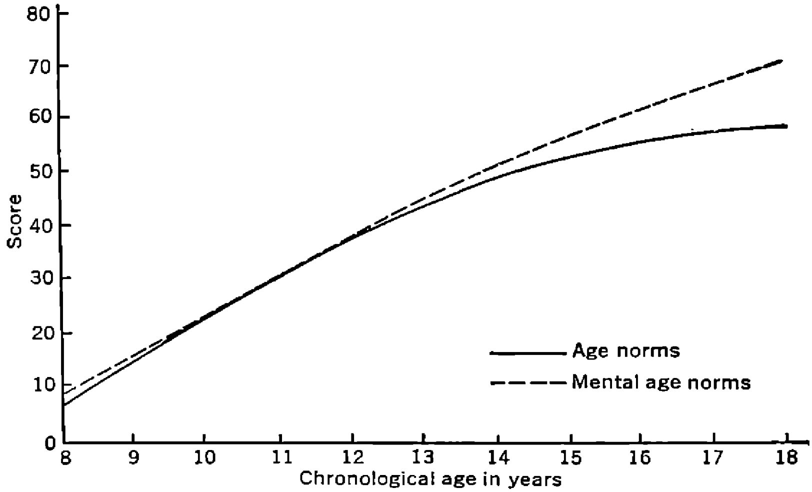


FIGURE 10. — Curves of mental growth from age norms and mental age norms

mental ages. All the interpretations one wishes to make with intelligence test scores can be easily made by using a table of norms such as is given for the "Terman Group Test of Mental Ability" on page 70. Indeed, the teacher or principal who wishes to use intelligently the results of an intelligence test given in his school will make comparisons of the scores of one pupil with others in his own school or with norms furnished by other schools, colleges, etc. Mental age as a concept, however, is very valuable as a means of expressing mental level. Mental age will be used subsequently in this book as synonymous with mental level, and reference to it as a concept does not imply that it should be used in the interpretation of pupils' scores in the high school.

Intelligence Quotient (I Q)

The intelligence quotient is found by dividing mental age by chronological age. Chronological age is the age we give when asked how old we are. $\frac{M\,A}{C\,A} = I\,Q$ is a convenient way of remembering how to find the I Q. For instance, if a 10-year-old boy is found to have a mental age of 12, then his I Q is $\frac{12}{10} = 1.20$. Usually, however, the decimal point is omitted and we write 120. Suppose another boy 10 years old has a mental age of 8, then his I Q is $\frac{8}{10} = 80$. The average 10-year-old boy has an average mental age of 10 and an I Q of 100; that is, the average I Q of people in general is 100. People who are dull or stupid have I Q's less than 100; bright or intelligent people have I Q's greater than 100. Half of the people have I Q's less than 100 and half greater than 100. Half of the people of the country have I Q's between 92 and 108. If you were to stand on a busy street corner and watch 100 people go by, about 50 of them would have I Q's between 92 and 108.

About 25 per cent of the people have I Q's less than 92. It is very seldom that one sees a person with an I Q less than 50 or so.

Terman states that persons with I Q's below 70 are definitely feeble-minded. A more cheerful side of this story is that about 25 per cent of people have I Q's greater than 108. These persons become the leaders in our society. They become our doctors, engineers, administrators, teachers, lawyers, ministers, business men, etc. Some of the children with high I Q's now in school are growing up to become the leaders in the next generation. We ought definitely to recognize that this is to be the case and pick out these children for special attention in the schools. These children with I Q's above 108 will probably all go to high school and many, if not most of them, will go to college.

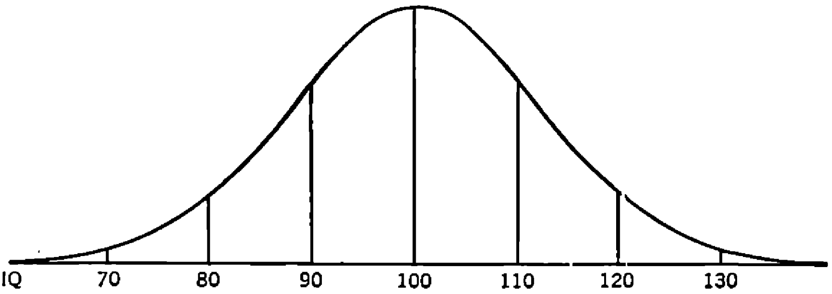


FIGURE 11. — Distribution of I Q's

Teachers in the elementary schools should be aware of this and should treat them accordingly. Whether such bright children become leaders in society, professional men and women with broad human interests, or whether they become mere moneymakers, unsocial in their attitudes, and working merely for private ends, depends largely on the type of education they receive. As Thorndike[1] says, "We cannot create intellect but we can prevent such a lamentable waste of it."

The distribution of I Q's is represented by our familiar probability curve. The majority of people have average intelligence, few people have low intelligence, and few people have high in-

[1] THORNDIKE, E. L. — *Educational Psychology*, vol. iii, p. 313.

telligence. The classical example of the distribution of I Q's is that given by Terman for 905 unselected children 5 to 14 years of age.[1]

There is much confusion in popular opinion regarding intelligence, due to the confusion concerning the underlying concepts of mental age and intelligence quotient. The concept of mental age is that of the general level of intelligence. As children grow older they are able to do more difficult and more complicated things. A child this year is not able to tell his name. Next year

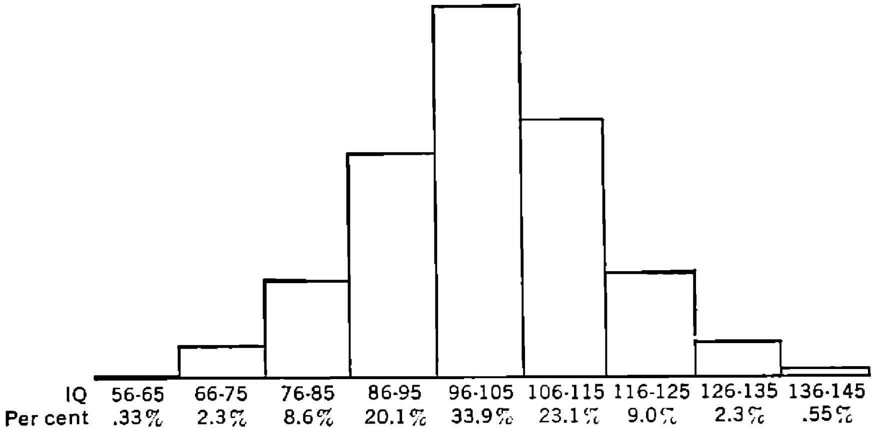


FIGURE 12. — Distribution of I Q's for 905 unselected children from five to fourteen years of age (from Terman)

he may be. A child this year is not able to solve a certain problem in arithmetic. Next year he may be. So the score on an intelligence test such as the "National Intelligence Test" corresponds more closely to the mental age than it does to the intelligence quotient. The intelligence quotient, on the other hand, refers to the intelligence of a person with respect to the intelligence of other people. The intelligence quotient of a child tells whether he is forward or backward for his age. The intelligence quotient does not change. A child of six has about the same intelligence

[1] Terman, *op. cit.*, p. 66.

quotient that he will have when he is fourteen. This is not absolutely true with our present tests because of test unreliability and because individuals fluctuate more or less, but enough experimental work has been done to enable us to say that theoretically the intelligence quotient remains constant all through life. Intelligence is determined before we are born.[1]

We must be careful in interpreting a teacher's statement about the brightness of a child. If a teacher says that a child is bright enough to be promoted, she is referring to mental age. If a teacher says that a child is bright and she knows that he will go to college, she is referring to the intelligence quotient. The mental age refers to the present level of intelligence. The intelligence quotient, since it remains practically constant throughout an individual's life, can be used to *predict* the limits of mental growth.

It was stated that mental age as a unit has several inadequacies for use in the high school. Since I Q is computed from the M A, the I Q suffers from the same inadequacies. The I Q also has difficulties of its own on the high school level which make it doubly objectionable as a unit. The denominator of the I Q formula is chronological age. Somewhere in the teens pupils begin to reach

[1] The experimental data concerning the constancy of the I Q has been reviewed by Rugg and Colloton (RUGG, H. O., and COLLOTON, C. — "Constancy of I Q as Determined by Retests," *Journal of Educational Psychology*, 12: 315–322, Sept., 1924.) The general method of testing the constancy of the I Q is to retest the same children after the interval of a year or more and note the change in I Q. These reviews report their summarized findings as follows: (1) The average difference between first and second findings of I Q is 5 points I Q. (2) Typical positive differences are less than 6 points in 50% of the cases and typical negative differences are less then 3 points in 50% of the cases. (3) The coefficient of correlation between first and second findings of the I Q have values for different investigators at .84, .93, .95, .94, and .85.

The reliability of the I Q "National Intelligence Test" is of the same order as the scores themselves (.90, .86) and the standard error of the I Q may be set at about 4.5 (4.3 and 4.9 on two different trials). This means that the I Q obtained from a group test has a reliability such that the true value of the I Q *on that test* will vary from an obtained value on the average by not more than 3 points.

the limit of their achievement on intelligence tests. Since the numerator of the formula reaches a maximum, the denominator must also be held constant if the I Q itself is to remain constant. Suppose that a boy at the age of 16 is credited with a mental age of 16. His I Q would be 100. At age 20 his I Q would have dropped to 80 ($\frac{16}{20}$), assuming no growth in mental age in the meantime. At age 40 his I Q would have dropped to 40 and would continue to drop throughout life. So far as is known the I Q remains constant throughout the years of growth, and in order to have the same meaning the I Q should keep the same value throughout all adult life. Hence somewhere the denominator in the ratio C A must reach a maximum and remain constant. Terman has set 16 years as the value to be used in the denominator of the I Q formula when determining the I Q for adults. This device is only an expedient — it has several glaring deficiencies. In the first place, in order to parallel the slowing up in the rate of mental growth, the denominator of the I Q formula should not stop abruptly but should begin slowing down at the age of 14, not reaching its maximum until three or four years later. As the I Q is now found, chronological age proceeds regularly until the age of 16 and then stops abruptly. This results in I Q's which are too low from about the ages of 14 to 18 when mental growth begins to slow down.

A second inadequacy is that it is not known definitely when mental growth stops. The data on the age at which mental growth ceases are scanty. We assert that mental growth ceases at some age after those ages in which it still exhibits some growth. Thorndike,[1] for instance, finds mental growth to persist during the whole high school period. Teagarten [2] sees "no reason for thinking that with average intelligence cessation of development

[1] Cf. "On the Improvement of Intelligence Scores."

[2] TEAGARTEN, F. M. — *A Study of the Upper Limits of the Development of Intelligence.*

occurs before $17\frac{1}{2}$ or 18 years of age." It seems certain that it stops so gradually that it is difficult to state just the age at which it ceases. But some value must be assumed for the denominator of the I Q formula for finding the I Q of adults.

A third inadequacy in finding the I Q is that it is probably true that not all persons reach mental maturity at the same age.[1] Consequently the same value should not be used in the denominator of the I Q formula in finding the I Q's of all adults. For persons with low I Q's the maximum value of the chronological age in the I Q formula may be 14 while for persons with high I Q's it may be 16.

Several paragraphs back the distinction between mental age and intelligence quotient was made. Let us call these two concepts "mental level" and "brightness." During the ages when mental growth is most rapid, mental level and brightness have two quite different meanings. As the adult state is reached these two concepts become less distinct, and for adults mental level and bright-

[1] There is considerable probability that mental growth does not stop at the same age for all individuals. The evidence seems to point to the existence of a positive correlation between I Q and the age at which mental growth stops. Doll * has published data showing that the growth of mental age of the feeble-minded stops at the age of 13. Stevenson Smith in a verbal report at the American Psychological Association in Washington December 31, 1924, stated that he found that certain of the races tested in Honolulu which showed low achievement on various intelligence tests gave evidence of reaching mental maturity at an earlier age than other races making higher intelligence test scores. Dearborn ("Intelligence Quotients of Adults and Related Problems," *Journal of Educational Research*, 6:307–325, Nov., 1922) believes that mental growth stops between the ages of 14 and 15, basing his belief on data including tests of children in continuation schools as well as in the regular public schools. The children in continuation schools would normally have dropped out of school and would not be included in tests made on regular school children. On the other hand, Thorndike (*Op. Cit.*) finds growth in mental ability throughout the high school period for children in high school — a group which becomes more and more highly selected for intelligence.

* Doll, E. A. — "The Average Mental Age of Adults," *Journal of Applied Psychology*, 3:317–328, Dec., 1919.

ness become identical. For this reason I Q as a unit has much less importance in the high school than it has in the elementary school. In the upper classes of the high school mental level is almost synonymous with brightness. In other words, age becomes an almost negligible factor in considering intelligence in the upper high school classes. It is a common practice to turn the scores on the "Terman Group Test of Mental Ability" into I Q's. Most people seem to think that they must have intelligence expressed in terms of I Q. This may be debatable in the 8th and 9th grades — it is certainly not true after that. As a matter of fact, for all of the purposes for which intelligence tests are used in high schools the I Q is quite unnecessary, — test scores may be safely used. *Mental level is the primary basis on which to section and test scores are a satisfactory measure of mental level in the high school. The I Q is not a primary basis on which to section in high school. Consequently the I Q is robbed of its main function and may be disregarded.* More extended discussion of this point is contained in Chapter XXIII.

There is more of a need for the I Q in guidance, but the following table adopted from data by Allen[1] shows that a classification according to mental level in high school also partially groups according to I Q. There is no corresponding relationship while children in the elementary school are growing.

Low down in the elementary school the correlation between mental age and I Q is not so high as exhibited in this table. The above table is for the class of elementary school graduates in Providence, R. I. As one ascends into high school, mental level and brightness continue to approximate a perfect correlation. Since guidance should be on the basis of brightness rather than mental level, there is more need for the I Q in this connection. Guidance which depends on I Q for its interpretation

[1] Allen, R. D. — Some Neglected Problems in Classifying High School Pupils," *Journal of Educational Research*, 11 : 351–358, May, 1925.

TABLE 83

DISTRIBUTION OF MENTAL AGES AND I Q'S IN THE GRADUATING ELEMENTARY SCHOOL CLASS, PROVIDENCE, R. I. (ADAPTED FROM ALLEN)

I Q	Mental Age							
	11–12	12–13	13–14	14–15	15–16	16–17	17–18	18–19
154–								1
145–153								2
136–144						2	12	2
127–135						22	21	
118–126					30	54	6	1
109–117			2	24	81	22	4	
100–108		9	144	199	62	5	1	
91–99		50	109	27	3			
84–90	2	69	40	4				
78–83	25	29	8					
73–77	18	9						
–72	8							
Totals	53	166	303	254	176	105	44	6

should be offered at the age of 12 or 13 or earlier, before all except very exceptional children have entered high school. By the time pupils have entered high school guidance should properly be of a more differential sort than that which depends on I Q.

The conclusion is that I Q is a unit that may be dispensed with in the high school. Test scores serve all purposes where a measure of intelligence is needed.

ACCOMPLISHMENT RATIO (A R)

The accomplishment ratio is the latest of our interpretative units. It is also very significant. It is found by dividing the

educational age by the mental age. The formula is $A\,R = \frac{E\,A}{M\,A}$.[1] The significance of this will be readily understood by referring to an example. Let us suppose we have a child with an M A of 10 years and an E A of 12 years. His A R is 120. What would you say of this child? One thing is probably true: the child is studious. It is probably also true that the child is forced through school by his parents. All in all, he is doing much better in his school work than most children of the same mental ability. Accordingly, the accomplishment ratio has been used as a measure of effort or motivation. Let us consider another child also with an M A of 10 years but with an E A of 8 years. This child has an accomplishment ratio of 80. We would say that he is not a studious child or has missed his early schooling or has trouble with the language or does not like school. His achievement is not what we expect of a child of his mental abilities. Again we see that the accomplishment ratio is a measure of effort or motivation or opportunity. The average accomplishment ratio is 100. Children who like school, who work hard, and who are given every incentive to learn, have A R's over 100. Children who dislike school, who are lazy, or who have not had the opportunity to learn, have A R's under 100.

Franzen,[2] one of the originators of the A R concept, at first believed that except for chance the A R never rises above 1.00. Franzen, for instance, put himself on record: "One's differences

[1] The A R may also be found by using the formula $A\,R = \frac{E\,Q}{I\,Q}$. The relationship of these two formulas may be seen from the following equations:

$$A\,R = \frac{E\,Q}{I\,Q} = \frac{\frac{E\,A}{C\,A}}{\frac{M\,A}{C\,A}} = \frac{E\,A}{M\,A}.$$

where *EA* stands for educational age and *EQ* stands for educational quotient.

[2] Franzen, R. H. — "The Accomplishment Quotient of School Marks in Terms of Individual Capacity," *Teachers College Record*, 21: 432–440, Nov., 1920.

when E Q is subtracted from I Q are always positive when they are large enough to be significant and small enough to seem spurious when they are negative. . . . It is safe, therefore, for practical use to assume that the optimum accomplishment is 1.00." This belief is dispelled by distributions of A R's as they are obtained from actual testing. Following is a distribution of A R obtained by using the "National Intelligence Test," Scale *B*, and the "Woody-McCall Arithmetic Tests."

TABLE 84

DISTRIBUTION OF ARITHMETIC A R DERIVED FROM THE NATIONAL INTELLIGENCE, SCALE *B*, AND THE WOODY-MCCALL ARITHMETIC TESTS

NATIONAL INTELLIGENCE TEST, SCALE *B*	F
135–139	2
130–134	2
125–129	12
120–124	9
115–119	16
110–114	30
105–109	29
100–104	34
95–99	21
90–94	18
85–89	17
80–84	3
75–79	2

From this distribution it may be seen that the A R goes well above 100. Indeed, it tends to form a normal curve of distribution centering about 100 as a mean, and in all respects is much like the I Q curve.

Bagley and Keith [1] have perpetuated this grotesque conception of the A R. They say:

[1] BAGLEY, W. C., and KEITH, J. A. H. — *An Introduction to Education*, p. 194, 1924.

Theoretically, it would seem that an accomplishment quotient (ratio) could never exceed 100, but there are on record cases of children whose intelligence as measured by the mental tests and recorded in I Q's is only moderate, but whose accomplishment has notably surpassed their assumed abilities. As one writer has expressed it, "So well have the teachers invested the intellectual capital of these children that the interest actually exceeds the principal." (Quotation from Cyril Burt: "Mental and Scholastic Tests.")

This conception of the A R might be called the "tank" conception. Intelligence is the tank to be filled with education; education is the liquid which teachers pour into the tank in order to fill it "to capacity" — and the A R is a measure of the per cent or degree to which the tank (intelligence) is filled with education. Such a conception of either intelligence or education does a grave injustice to both. Intelligence is not a *tabula rasa* to be scrawled over with writing. Intelligence is rather to be thought of as an *average* or general ability — actual acquired learned ability — acquired at school, home, in the street, playground, churches, and movies. Achievement is to be thought of as special ability along some one line. Naturally this special ability may surpass or be inferior to general average ability (intelligence) and the A R will accordingly be above or below 1.00.

The A R varies much in reliability, depending both on the reliability of the original tests and on the likeness or difference of material in the intelligence and achievement tests. The reliability of the reading A R as found from the "Thorndike-McCall Reading Scale" and the "National Intelligence Test," Scale *A*, is .34 and from the "National Intelligence Test," Scale *B*, is .23. These reliability coefficients correspond to standard errors of 6.8 and 7.3 in A R units. These low reliability coefficients may be partially attributed to the dependence of intelligence tests on reading ability, so that in part the intelligence tests and the reading tests are a measure of the same function. Differences between these tests, then, ought to be nearly random. The cor-

responding arithmetic A R's, using the "Woody-McCall Mixed Fundamentals Arithmetic Scale" and the "National Intelligence Tests," is .60 and .49. The corresponding standard errors are 6.4 and 6.2.

Educators have grasped at the accomplishment ratio as a very valuable unit. It has been suggested as a mark for the evaluation of school work. Used as a mark it rates the work of every pupil according to his merits. The bright child must do more than the dull pupil to get the same accomplishment ratio. As it is now in school the race or competition is often artificial. We offer prizes as though any one could win them, whereas we know perfectly well they will be won by the bright pupils. The accomplishment ratio as a mark sets a goal according to the pupil's ability and gives him credit according to whether he fulfills the expectations.

The accomplishment ratio has also been suggested as a measure of a teacher's efficiency. If the teacher is liked by the pupils and stimulates them to do their best work, she will produce results surpassing those obtained by the average teacher working with a group of the same ability. The average accomplishment ratio of such a class would be over 100. If the teacher is harsh and tyrannical, breeding hatred in the class and driving instead of leading the pupils in their school work, she will obtain results less than those obtained by the average teacher working with a group of the same ability. The average accomplishment ratio of such a class would be under 100. Not only two teachers, but two schools, might be compared by making use of the accomplishment ratio. This explains what was said in another chapter with regard to a teacher's not being judged by the degree to which she brings her class up to a grade norm. What a teacher does for the class depends, in the first place, on the nature of the class with which she has to deal. The teacher with a C section is not expected to produce the same results as the teacher with an A section. Likewise the prin-

cipal of a school situated in a select residential section is expected to produce better results in his school than a principal in a school less fortunately situated.

What are the causes of high or low accomplishment ratio? The main cause is the factor of opportunity. If a child is in a section or grade too low for his ability, he does not have the opportunity to learn the things learned by other children of his mental ability. This affects bright children more than it does dull children, for the bright children are the ones who have not been sent forward fast enough in school. Likewise children who are in a grade too difficult for their mental ability are likely to have A R's over 100. These children are pushed and are usually among the dullest children. All told, dull children work harder than bright children in school. By rights, just the opposite should be the case. The bright children of whom we expect so much should be the ones whom we should especially encourage to make the most of their mental ability. A second cause for high or low A R is the encouragement or lack of encouragement which is offered the pupil in school and at home.

The accomplishment ratio enables us to see more clearly the value of life. Rewards are usually given in this world for high educational age or its equivalent, meaning by this any kind of success or achievement which the world thinks valuable. In the past it has been thought that high achievement is potentially possible for every one. The writers of the Declaration of Independence stated that all men are created free and equal. We are thankful that to-day we are politically free, but psychology has taught us that we are not potentially equal, and as a general rule the prizes in life have gone to those with the best hereditary endowment. The concept of the accomplishment ratio shows us a new way to judge people. Every man is to be judged by the degree to which his achievement measures up to his capacity. It is precisely the sort of judgment which is exhibited in the

parable of the talents. So in school, teachers should first find out the ability of their children and then should judge their work in the light of this ability.

Naturally, since we have discarded all ages and quotients, the accomplishment ratio as such cannot be found or used in the high school. But the concept of effort or motivation or studiousness which the accomplishment ratio represents is one of our most valuable and effective tools of thought. In Chapter XXIV, under the topic an "index of studiousness," a makeshift unit serving the same purpose as the accomplishment ratio will be described. This makeshift unit I have termed "sigma difference." In a separate monograph[1] will be found ability standards for nine standardized tests in high school subjects, which permits a more accurate determination of studiousness. This task of providing ability standards is a vital one for the greater usefulness of standardized tests in the high school at the present time. By ability standards is meant standards corresponding to various intelligence levels. With such standards set for different levels of ability it is possible to measure the degree to which such standards are or are not surpassed.

The discussion has ended with no usable derived units for use in the high school. Mental and educational age, intelligence and educational quotients, and the accomplishment ratio, although standing for very real concepts, are not suitable for computation on the high school level. Mental age reduces itself to an absurdity or at best to a mere arbitrary unit, I Q and M A tend to coincide as maturity is reached, and the A R, depending on these other units, is equally faulty when actually computed. Mere test scores are all that is left. As stated at the opening of this chapter, test scores contain all the information that a test has to yield. Nothing additional is gained by turning the scores into derived units. Once one understands how to interpret a test score he

[1] To be published by the Bureau of Publications, Teachers College.

knows all that is necessary to make the best use of that test in the high school. Educational measurements have been obscured by an overdevelopment of the systems of derived scores. This chapter has suggested reasons why the present derived scores are unsuited for use in the high school.

REFERENCES

ALLEN, R. D. — "Some Neglected Problems in Classifying High School Pupils," *Journal of Educational Research*, 11 : 351–358, May, 1925.

BAGLEY, W. C., and KEITH, J. A. H. — *An Introduction to Education*, ch. viii. The Macmillan Company, New York; 400 pp., 1924.

BEESON, M. F., and TOPE, R. E. — "The Educational and Accomplishment Quotients as an Aid in the Classification of Pupils," *Journal of Educational Research*, 9 : 281–292, April, 1924.

BROOKS, F. D. — *Changes in Mental Traits with Age; Determined by Individual Retests*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 116, Teachers College, Columbia University, New York; 88 pp., 1921.

— "Rate of Mental Growth, Ages Nine to Fifteen," *Journal of Educational Psychology*, 12 : 502–510, Dec., 1921.

BROOKS, S. S. — *Improving Schools by Standardized Tests*, ch. iv. Houghton Mifflin Company, Boston; 273 pp., 1922.

DEARBORN, W. F. — "Intelligence Quotients of Adults and Related Problems," *Journal of Educational Research*, 6 : 307–325, Nov., 1922.

DICKSON, V. E. — *Mental Tests and the Classroom Teacher*, ch. i, iv, v. World Book Company, Yonkers-on-Hudson, New York; 231 pp., 1923.

DOLL, E. A. — "The Average Mental Age of Adults," *Journal of Applied Psychology*, 3 : 317–328, Dec., 1919.

FRANZEN, R. — "The Accomplishment Quotient of School Marks in Terms of Individual Capacity," *Teachers College Record*, 21 : 432–440, Nov., 1920.

— *The Accomplishment Ratio*. *Ref*. Ch. 13.

GARRISON, S. C. — "Additional Retests by Means of the Stanford Revision of the Binet-Simon Tests," *Journal of Educational Psychology*, 13 : 307–312, May, 1922.

GREGORY, C. A. — *Fundamentals of Educational Measurement*, ch. vi. *Ref*. Ch. 2.

KELLEY, T. L. — "A New Method for Determining the Significance of Differences in Intelligence and Achievement Scores," *Journal of Educational Psychology*, 14 : 321–333, Sept., 1923.

McCALL, W. A. — *How to Measure in Education*, chs. ii, v. *Ref.* Ch. 3.

MONROE, W. S. — *Introduction to the Theory of Mental Measurements*, ch. vii. *Ref.* Ch. 3.

— "Illinois Examination," *Bulletin no. 6 of the Bureau of Educational Research*, University of Illinois; 70 pp., 1921.

MONROE, W. S., and BUCKINGHAM, B. R. — *Teacher's Handbook*, Illinois Examination. *Ref.* Ch. 13.

MONROE, W. S., DEVOSS, J. C., and KELLY, F. J. — *Educational Tests and Measurements*, Ch. x. *Ref.* Ch. 2.

MURDOCK, K. — "The Accomplishment Quotient — Finding and Using It," *Teachers College Record*, 23: 229–239, May, 1922.

OTIS, A. S. — *Statistical Method in Educational Measurement*, chs. xii, xiii, xiv. *Ref.* Ch. 12.

PINTNER, R., and MARSHALL, H. — "A Combined Mental-Educational Survey," *Journal of Educational Psychology*, 12 : 32–43, Jan., 1921.

— "Results of the Combined Mental-Educational Survey Tests," *Journal of Educational Psychology*, 12 : 82–91, Feb., 1921.

POULL, L. E. — "Constancy of Intelligence Quotient in Mental Defectives," *Journal of Educational Psychology*, 12 : 323–324, Sept., 1921.

RUCH, G. M. — "The Achievement Quotient Technique," *Journal of Educational Psychology*, 14 : 334–343, Sept., 1923.

RUGG, H. O., and COLLOTON, C. — "Constancy of Intelligence Quotient as Determined by Retests," *Journal of Educational Psychology*, 12 : 315–322, Sept., 1921.

RUGG, L. S. — "Retests and the Constancy of the Intelligence Quotient," *Journal of Educational Psychology*, 16 : 341–343, May, 1925.

SMITH, S. — *Race Differences in Hawaii.* Reported at the Thirty-third Annual Meeting of the American Psychological Association, Washington, District of Columbia, Dec. 31, 1924.

STEBBINS, R., and PECHSTEIN, L. A. — "Quotients I, E, and A," *Journal of Educational Psychology*, 13 : 385–398, Oct., 1922.

SYMONDS, P. M. — "The Accuracy of Certain Standard Tests for School Sectioning and Marking," *Journal of Educational Psychology*, 15 : 423–432, Oct., 1924.

TEAGARTEN, F. M. — "The Constancy of the I Q Again," *Journal of Educational Psychology*, 13 : 366–372, Sept., 1922.

TEAGARTEN, L. M. — *A Study of the Upper Limits of the Development of Intelligence*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 156. Teachers College, Columbia University, New York; 112 pp., 1924.

TERMAN, L. M. — *The Measurement of Intelligence*, chs. iii, viii. Houghton Mifflin Company, Boston; 362 pp., 1916.

— *Intelligence Tests and School Reorganization*, chs. ii, iv. World Book Company, Yonkers-on-Hudson, New York; 111 pp., 1923.

— *The Intelligence of School Children.* Houghton Mifflin Company and others, Boston; 317 pp., 1919.

THORNDIKE, E. L. — "On the Improvement of Intelligence Scores from Fourteen to Eighteen." *Ref.* Ch. 13.

TOOPS, H. A., and SYMONDS, P. M. — "What Shall We Expect of the A Q?" *Journal of Educational Psychology*, 13 : 513–528, Dec., 1922; 14 : 27–37, June, 1923.

TRABUE, M. R. — *Measuring Results in Education*, ch. xix. *Ref.* Ch. 4.

WHIPPLE, G. M. — "Intelligence Tests and Their Uses." *Ref.* Ch. 4.

CHAPTER XVI

THE MEASUREMENT OF CONDUCT

Rating

The measurement of conduct — of what a person *will* do under certain circumstances — has given considerable difficulty to educational psychologists. Nearly all of our mental and educational tests are tests of ability — what a person *can* do. All tests presuppose a test situation and one element in the test situation is that the person is striving to do his best. Under normal conditions a test usually does show what a person *can* do when motivated. But a test of conduct implies that the person is unaware of being tested and is reacting normally to typical situations as they present themselves.

Conduct is usually considered qualitatively or descriptively. "John is prompt, tidy, trustworthy, and a good sport," succinctly describes numerous special conduct tendencies of John more or less accurately. The probabilities are that John is not always prompt, nor is he prompt in all situations, wherefore the above statement needs qualification. The above descriptive language is on an "all or none" basis, but conduct is not dichotomous in such a manner.

The measurement of conduct may take two forms: (1) the judgment of conduct according to a more refined technique than that just described or (2) the attainment of some record of the conduct which permits the making of an objective measure.

The first of these resources is psychology's ultimate resort in case more objective methods are not available. One can always fall back upon perception and judgment. Continued experi-

mentation with the making of judgments has led to some knowledge of the pitfalls to be avoided and has led to the establishment of a technique by means of which the judgment may most economically and objectively be recorded.

A first refinement of the rating of conduct over and above that employed by the characterization in our ordinary English is the placing of the individual to be measured somewhere along a scale on a trait according to the judgment of the rater. In making this judgment one rater tries to think of the *average* status of the individual being measured in the trait under consideration, having in mind *no special instance* when the individual exhibited the trait in question. To be more specific, suppose that the rater is passing his judgment as to the honesty of an individual. He does not qualify him as being merely honest or dishonest, but, thinking of a scale running from the utmost dishonesty through varying degrees of less than perfect rectitude to perfect uprightness, tries to place the individual somewhere along that scale according to his judgment. Later it will be shown that this method, even with all possible refinement and external aid, is relatively unreliable, so that a single person's judgment of an individual in any one trait is not usually to be depended upon. Only when several acquaintances pass their independent judgments and these judgments are pooled is a reliable rating achieved.

Instead of judging traits one may rate the expected response in a particular situation. This may be called conduct rating as opposed to trait rating. That is, instead of rating an individual as to his honesty one might pass judgment as to whether or not he would cheat in an examination. This conduct rating may be of the simple dichotomous kind where a judgment is passed "He will not cheat in an examination." More analytically, each conduct situation is not merely a question of plus or minus — there is always a probability, varying from 0 to 1, of the response being made. On a five-point scale this probability might be expressed

as never, seldom, sometimes, often, always. This probability at any given moment may take any value. The case is analogous to measurement in arithmetic. In the testing of a single addition combination — 5 + 7 — there is always a probability less than 1 that the correct response will be given. Sometimes an error will be made once in every ten trials, sometimes only once in a hundred trials. The error, if it occurs only so infrequently, would be explained by perseveration or anticipation or some other lapse where a "wrong" stimulus or a "wrong" set momentarily obtained the ascendency. Precisely in conduct, some subtle element in the situation — either inner or outer — may cause aberration in the response, and these slight aberrations in the response are the cause of the less-than-1 probability of a uniform response. In conduct responses there are usually social elements in the situation which furnish many prepotent stimuli to response irregularities. For instance, a pupil may not cheat in an examination once in a hundred times when the teacher is in the room, but may cheat if the teacher steps out. This pupil, however, is relatively honest in examinations as compared with the pupil who cheats on every possible opportunity with the teacher's back turned for an instant.

The reality of a trait has often been questioned. Is an individual always exactly honest as regards playing games or in paying his fare on a street car? When one stops to think of just how the rating on some trait is made, only vague recollections are brought to mind. Scanty evidence at best is all that we can recall when we are asked concerning a person's honesty. One perhaps remembers a laxity in an individual in returning some overchange given back by a storekeeper, or perhaps the telling of an event was misrepresented, but unless certain events stand out clearly in memory, one tends to class all people as of even rectitude and trustworthiness. To give an adequate rating one should have noticed the individual carefully in many situations where the

trait might be exhibited. For instance, if it is a question of honesty, one might observe and give answers to such questions as:

1. Does he return borrowed money promptly?
2. Does he return money when he receives too much in change?
3. In traveling does he pay his fare without any attempt to dodge?
4. Does he return found articles?
5. Does he keep a promise?
6. Does he obey rules?
7. Does he play games honestly?
8. Does he refrain from cheating in an examination?
9. Does he recite without attempting to bluff?
10. Does he do his work independently?

So there is a relationship between trait rating and conduct rating. The element of behavior is the reaction in a certain situation, and one's primary judgments are always based on his observations of actual conduct. One may collect his impressions of a group of related conduct trends into what is called a trait. It seems certain that trait judgment ultimately resolves itself into the observation of separate elements of conduct, each a phase or exhibition of the trait in question. It remains for experimental verification to prove that the average of the judgments on the above 10 points for any individual is identical with the rating of the individual in honesty in general. Although close, probably they are never identical. Of the two, however, one should choose the summation of answers to the ten questions as a measure of honesty rather than a rating of honesty itself. For one knows that the answers to the ten questions represent a fair sampling of the field of honesty, whereas a blanket rating may rest on evidence of two or three scattered and partial observations.

Rating Methods

Rating is not new. Expressive adjectives have always been used to express judgments of personal characteristics. The first mention that I can find of a scale of several items by which judgments may be made was that constructed by Galton in 1883 in a section on Mental Imagery in his *Inquiry into Human Faculty and Its Development.* This is a scale which may be used for rating the quality of one's mental imagery. I have reproduced the scale because it remains to-day a paragon of its kind.

Scale of Mental Imagery — Imagery of a Breakfast Table

Highest. — Brilliant, distinct never blotchy.

First Suboctile. — The image once seen is perfectly clear and bright.

First Octile. — I can see my breakfast table or any equally familiar thing with my mind's eye quite as well in all particulars as I can if the reality is before me.

First Quartile. — Fairly clear; illumination of actual scene is fairly represented. Well defined. Parts do not obtrude themselves, but attention has to be directed to different points in succession to call up the whole.

Middlemost. — Fairly clear. Brightness probably at least from one half to two thirds of the original. Definition varies very much, one or two objects being much more distinct than the others, but the latter come out clearly if attention be paid to them.

Last Quartile. — Dim, not comparable to the actual scene. I have to think separately of the several things on the table to bring them clearly before the mind's eye, and when I think of some things the others fade away in confusion.

Last Octile. — Dim and not comparable in brightness to the real scene. Badly defined, with blotches of light; very incomplete; very little of one object is seen at one time.

Last Suboctile. — I am very rarely able to recall any object whatever with any sort of distinctness. Very occasionally an object or image will recall itself, but even then it is more like a generalized image than an individual one. I seem to be almost destitute of visualizing power as under control.

Lowest. — My powers are zero. To my consciousness there is almost no association of memory with objective visual impressions. I recollect the table but do not see it.

A sample of the modern scale is the following taken from Thorndike's *Psychology of Algebra.*[1]

Scale for Measuring the Importance of Technical Terms in Algebra

10. Essential for economical and efficient learning.
8. Very useful.
6. Probably worth learning.
4. Probably not worth learning.
2. The time is much better spent otherwise.
0. Of no value at all for the learning of algebra.

Use 1 when in doubt whether to use 0 or 2;
Use 3 when in doubt whether to use 2 or 4; etc.

These scales represent the method of expressing one's judgment by rating. Another method is that of ranking.

RATING *vs.* RANKING

Although rating scales are generally preferred, ranking is also sometimes used and has generally been considered the more reliable. By rating is meant the assigning of a position on a scale of already defined steps. Rating is the method used in assigning a value to a composition on the Hillegas Scale or to a sample of handwriting on the Thorndike Scale. It is the method one uses in the assertion that a child is in excellent, good, fair, or poor health, the adjectives used roughly in defining the steps on a four-point scale. By ranking is meant the placing of individuals in order as in arranging a gymnasium class in order of height or in sorting the cards in a pack in order from ace to king.

It is commonly thought that ranking is more reliable than rating because it demands a closer comparison. Allport [2] says, "The ranking method, available only in rating an actual group all of whom are known to each rater, is *more accurate.* Com-

[1] P. 264.

[2] ALLPORT, F. H. — *Social Psychology*, p. 128, 1924.

parisons are more concrete and definite than in the scoring method." In rating, comparison is made with roughly defined steps or classes in the scale and the individual is placed in the nearest fitting class according to the opinion of the rater. But in ranking, one individual is compared directly with another.

In this connection Knight [1] makes another pertinent comparison between the methods of rating and ranking:

> Judges who are presumably equally competent rate the same person as almost wholly honest, half honest, rather dishonest, or an out-and-out cheat. This disagreement is the product of legitimate variations in knowledge about the person and of varying standards of honesty held by the judges themselves. The way to avoid this pitfall is not to use a rating involving concepts of virtues, but a system of ranking persons according to their relative merits. The question should be not "How honest is he?" but "Is he more or less honest than this person or that?"

Knight's objection to rating scales is not wholly an objection to their differences in reliability — he is objecting to the tendency to rank high or low on a rating scale as a whole. One teacher might give a class an average rating of 4 on a rating scale in honesty and another teacher give the same class a rating of 2 and yet the two might correlate perfectly. In using the ranking method we can transmute the ranks into sigma indices by means of tables as given by Chu [2] or into percentiles by means of tables as given by Haught [3] and can assume that the means of different groups are the same, an assumption which probably is nearer the truth in most cases than taking the means to be as different as may often be found on rating scales. On the other hand, many testify that rating is more pleasing than ranking. Ranking grows con-

[1] Knight, F. B. — "Analysis of Teaching and Teachers," *Journal of Educational Research*, 10: 224, Oct., 1924.

[2] Chu, J. P. — *Chinese Students in America: Qualities Associated with their Success.*

[3] Haught, F. B. — "A Scheme for Combining Incomplete Rankings," *Journal of Applied Psychology*, 7: 168–172, June, 1923.

fusing when the groups are large and the effort to make fine distinctions between individuals becomes irksome. Experimentation shows that the fine distinctions simply are not made. Symonds [1] cites data to show that rating and ranking are equally reliable, and so there is no reason for preference of one or the other method on that score. A combination of the two, which keeps the merits of both and which seems superior to the use of either the rating or ranking method singly, will shortly be described. A question to be decided before any rating is to be done is that of the number of classes in which to group. The following table taken from Boyce [2] shows that there is much variance in actual practice as to the number of classes a rating scale may have.

TABLE 85

DISTRIBUTION SHOWING NUMBER OF CLASSES IN 54 RATING SCALES (FROM BOYCE)

NUMBER OF CLASSES IN RATING SCALE	FREQUENCY
2	2
3	11
4	24
5	12
6	4
7	1
	54

Naturally one wishes in any measurement to use as fine a scale as possible, but just as in physical measurements it is useless to use

[1] SYMONDS, P. M. — "Notes on Rating," *Journal of Applied Psychology*, 9 : 188–195, June, 1925.

[2] BOYCE, A. C. —"Methods for Measuring Teachers' Efficiency," being Part II of the *Fourteenth Yearbook of the National Society for the Study of Education*, p. 20, 1915. By permission of the Public School Publishing Company.

a scale finer than the limits of one's eyesight or finer than the definiteness of the object to be measured, so in psychological ratings it is useless to use a scale finer than the judge's ability to discriminate or the definiteness of the trait in the subject. The degree of fineness to which this discrimination can be made in psychological ratings is determined by the coefficient of reliability. One has to choose how much of a loss in reliability he will permit in order to make easier the rating. Symonds presents arguments showing that in constructing scales for rating traits of personality the optimum number of class intervals is 7. Rating with a scale of more than this number of classes demands a discrimination which does not yield an increase in reliability sufficiently great to make the increase worth while according to an arbitrary definition of such worthwhileness. Likewise, according to our definition, a rating scale with a fewer number of classes suffers from a loss of reliability greater than is allowed, due to the coarseness of grouping.

This refers to the rating of human qualities or characteristics by reasonably competent and willing judges. If the qualities are obscure and abstruse or if the raters are untrained or noticeably careless, less than seven classes are sufficient. On the other hand, there are many things which may be judged more objectively than human characteristics, and if they are to be rated, more than seven classes in the rating scale should be used.

Rating may be most simply and straightforwardly accomplished by placing opposite the name of a person a number signifying his rating. First, there must be the description, defining steps on the scale—then these scale values may be used directly in the rating. But such rating is extremely subjective.

THE MAN-TO-MAN SCALE

"The first innovation and one promise of real progress (toward objectivity of rating) was the man-to-man comparison scale,"

says Rugg.[1] This was the product of a seminar discussion at the Carnegie Institute of Technology — a class conducted by Professor Walter Dill Scott. The suggestion was applied in the development of rating scales for employees in industry. Measurement, ran the argument, implied comparison with a scale. The conclusion followed — make your "rating scale" a scale of human beings. The increments — the unit distances on your scale — will be the distances or differences between "scale men." How select these scale men? First, by choosing "the best man you ever knew" and writing his name down at the top point on your rating scale. Second, by selecting "the poorest man you ever knew" and using him at the foot of the scale. Similarly you choose an "average man"; a "better-than-average man," and a "poorer-than-average man," making a five "point" scale. You give numbers to the five men, like 15, 12, 9, 6, and 3, and your "scale" is complete.

Rating was simple on this scale. You simply compared your man with these five men. Was he as good as the best man, — say, Richardson? No. Was he better than Johnson, the poorest man? Yes, — by far. Was he superior to Bankson, the "average" man? Well, hardly. And so it was until he was finally placed on the scale and given a score.

The man-to-man scale is not used to any great extent to-day, notwithstanding the impetus which its use received in the war. Its advantages are probably more apparent than real. The same close comparison between individuals may be achieved in other ways, one of which is described on page 345.

[1] RUGG, H. O. — "Is the Rating of Human Character Practicable?" *Journal of Educational Psychology*, 12 : 425–438, Nov., 1921; 485–501, Dec., 1921; 13 : 30–42, Jan., 1922; 81–93, Feb., 1922.

THE ARMY RATING SCALE

I.	PHYSICAL QUALITIES Physique, bearing, neatness, voice, energy, endurance. Consider how he impresses his command in these respects.	Highest . . . . . 15 High . . . . . 12 Middle . . . . . 9 Low . . . . . 6 Lowest 3
II.	INTELLIGENCE Accuracy, ease in learning; ability to grasp quickly the point of view of the commanding officer, to issue clear and intelligent orders, to estimate a new situation, and to arrive at a sensible decision in a crisis.	Highest . 15 High . 12 Middle . 9 Low . . 6 Lowest . 3
III.	LEADERSHIP Initiative, force, self-reliance, decisiveness, tact, ability to inspire men and to command their obedience, loyalty, and coöperation.	Highest . . . 15 High . . . 12 Middle . . . . . 9 Low . . . . . 6 Lowest 3
IV.	PERSONAL QUALITIES Industry, dependability, loyalty; readiness to shoulder responsibility for his own acts; freedom from conceit and selfishness; readiness and ability to coöperate.	Highest . 15 High 12 Middle . 9 Low . . 6 Lowest . 3
V.	GENERAL VALUE TO THE SERVICE Professional knowledge, skill, and experience; success as an administrator and instructor, ability to get results.	Highest . . . . . 15 High. . . . . 12 Middle . . . . . 9 Low . . . . . 6 Lowest 3

THE GRAPHIC RATING SCALE

In the graphic rating scale the rating is made by placing a cross somewhere along a line representing the scale of the function being measured. Two rating charts employing the graphic principle are reproduced here.

Rating Chart A

Pupil's Name..........................Date......................
School..............................Grade.....................
Rated by...

Directions for Using the Rating Chart

1. Let these ratings represent your own judgment. Do not confer with any one in making them.
2. In each trait or characteristic named below compare this pupil with the average pupil of the same age.
3. In any particular trait disregard every other trait except that one. Many ratings are rendered valueless because the rater allows himself to be influenced by a generally favorable or unfavorable impression which he has formed of the person rated. Do not rate a pupil high on all traits simply because he is exceptional in some. Children are often very high in some traits and low in others.
4. Place a cross somewhere on the line running from "very high" to "very low" to indicate this child's standing in each quality. You may place your cross at *any point* on the line. It is not necessary to locate it at any of the division points or above any descriptive phrase.
5. Do not study too long over any one trait. Give for each the best judgment you can and go on to the next.
6. Give a rating for every trait.
7. The ratings will be held strictly confidential.

Health — Is he generally healthy and vigorous?

Bad	Poor	Average	Good	Excellent

Leadership — Does he take the lead in school affairs or does he follow others?

Always follows others	Rather tends to follow	Average	Rather tends to be a leader	Masterly, not easily influenced

The usual method with such a scale is to go through the traits for one individual and then turn the sheet to the next individual. Even though instructed to rate all individuals on one trait before turning to the next trait, no direct comparisons can be made since

only one page is visible at a time. To make use of the close comparison which is demanded in ranking, rating scheme B is recommended:

Rating Chart B

Rated by..........................Date..........................
School..........................Grade..........................

Directions for Using the Rating Chart

1. Let these ratings represent your own judgment. Do not confer with any one in making them.
2. In each trait or characteristic named below compare this pupil with the average pupil of the same age.
3. In this particular trait disregard every other trait except this one. Many ratings are rendered valueless because the rater allows himself to be influenced by a generally favorable or unfavorable impression which he has formed of the person rated. Do not rate a pupil high on this trait simply because he is exceptional in some. Children are often very high in some traits and low in others.
4. Place a cross in one of the compartments running from "very high" to "very low" to indicate each child's standing in this quality.
5. Do not study too long over any one child. Give for each child the best judgment you can and go on to the next.
6. Give a rating for each child.
7. The ratings will be held strictly confidential.
8. Try to let the percentages guide you as to the number of crosses to fill in each compartment.

Trait: HEALTH

Is he generally healthy or vigorous?

Pupil	4% Very Bad	11% Bad	21% Poor	28% Average	21% Good	11% Very Good	4% Excellent
Charles . . .							
William . .							
George . . .							
etc.							

Scale B is to be generally preferred because it keeps the desirable features of the graphic rating chart and also permits the close comparison of ranking. To be sure, the ratings for any one individual are spread over several sheets, but as in any group test the ratings for the several tests for any individual may easily be brought together when the numerical values of the graphic rating are measured.

Freyd [1] lists as advantages of the graphic rating scale:

It is simply and easily grasped.
It is interesting and requires little motivation of the rater.
It is quickly filled out.
It is simply and easily scored.
It frees the rater from direct quantitative terms.
The descriptive terms aid the rater in that they make the various degrees of the trait more concrete.

It is also claimed that the rater is enabled to make his discrimination as fine as he cares. The fineness of the scoring may be altered at will by means of a scoring stencil, and may yield scores of from 1 to 5, or from 1 to 100. This fineness of discrimination is somewhat illusory. With more than 7 classes on the scale the finer discriminations in rating human traits simply are not made, and whatever additional increment in discrimination there may be affords only a small gain in reliability.

Freyd [2] also gives a list of "rules" to be followed in constructing a graphic rating scale from which the following are selections.

Define the trait on which you wish ratings. It will often be found that what one considered a single trait is in reality composed of a number of well-defined separate traits. Define the trait in terms of what the individual actually does, for the more concretely the trait is expressed, the greater is the expectation that raters will agree.

Decide on the extremes of the trait. It is frequently the case that one extreme of a scale may have several opposites.

[1] Freyd, M. — "The Graphic Rating Scale," *Journal of Educational Psychology*, 14: 83–102, Feb., 1923. [2] *Op. cit.*

It will be found good practice to introduce every scale with a question, to which the rating furnishes the answer; for instance, the question "How tactful is he?" or "Is he tactful or careless?" may be answered by checking on the rating line.

The line should not be much more than five inches in length, so that it may be grasped as a unit.

The end phrases of the scale should not be so extremely worded as never to be employed.

The phrase descriptive of the neutral or average degree of the trait should be in the center of the scale.

Only universally understood phrases should be used. Slang is effective if there is no doubt as to its meaning.

Such terms as *average, very, extremely, excellent, good, fair,* or *poor,* should be used sparingly. Use in their place adjectives which in themselves express the varying degrees of the trait. In place of *extremely neat* one might say *fastidious,* or in place of *very careless in dress* one might say *slovenly.*

The descriptive phrases should be short and to the point. The favorable extremes should be alternated so as to do away with a motor tendency to check at one margin of the page.

There is a tendency of raters to rate constantly too high or too low. Scale B on page 345 shows one method of avoiding this issue by suggesting the percentages that would lead to a normal distribution for any group. Hayes and Paterson have suggested the method of dividing the distribution of the ratings given by each supervisor into five classes on a percentage basis. Each total rating is thus converted into a final letter rating. It is evident that the final ratings remove the error arising from differences of standards of judgment, while they preserve all real disagreements.

Validity of Ratings

Ratings are ultimate. One can never get behind judgments for otherwise intangible traits or qualities. Ratings ought, therefore, to be perfectly valid. Unfortunately they are not, because human judgments are fallible. Human judgment not only tends to make errors, but constant errors. That is, oftentimes our judg-

ments are warped or biased. The lack of perfect validity shows itself in three ways.

HALO EFFECT

For some time users of rating devices have noticed that ratings of human traits are inaccurate because the rater's judgment of the trait in a certain individual is influenced by his general impression of that individual. So far as I am able to determine, Webb,[1] in his excellent and too little read monograph *Character and Intelligence*, was the first to notice this halo effect in 1915. "Let us suppose, for instance," he says, "that the observers, in estimating the intelligence qualities, are biased in the direction of marking subjects who possess other desirable qualities too highly and vice versa." In 1920 Thorndike[2] noticed this phenomenon in this country and used the term "halo" in identifying it. At that time he said, "The magnitude of the constant error of the halo, as we have called it, also seems surprisingly large, though we lack objective criteria by which to determine its exact size."

Rugg[3] sums up: "The suggestion comes insistently, however, that one of the most potent influences working against accurate estimates of character is the prevalence of just such general attitudes toward our associates and subordinates."

EFFECT OF ACQUAINTANCE

Knight[4] showed that intimate acquaintance of the rater with the individual being rated tends, strangely enough, to make the ratings less accurate.

It is commonly assumed that length of acquaintance increases the accuracy and reliability of the judgment. If the varying lengths of time and acquaint-

[1] Webb, E. — *Character and Intelligence.*

[2] Thorndike, E. L. — "A Constant Error in Psychological Ratings," *Journal of Applied Psychology*, 4: 25–29, Mar., 1920. [3] *Op. cit.*

[4] Knight, F. B. — "The Effect of the Acquaintance Factor upon Personal Judgments," *Journal of Educational Psychology*, 14: 129–142, Mar., 1923.

ance are one minute, two hours, three weeks, there may be a close direct relationship between worth of estimate and length of acquaintance. When varying lengths of acquaintance is a matter of years, we find that knowing the applicant too long apparently decreases the critical value of the judgments. . . . The factor of acquaintance, then, operates to make ratings more lenient, *i.e.*, increases the over-rating, and to make ratings less critical and less analytical, *i.e.*, increases the influence of the halo of general estimate. It is in the direction of truth to discount the ratings of judges when acquaintance has been long. In a way it is literally true to say of a judge's estimate: "His judgment is of doubtful validity *because* he has known his man *too long*."

Shen,[1] in a more critical study, arrives at the same conclusion. "Our best friends occupy such a favorable position in our view that whenever we compare them with others, we are comparing, as of the four-shaped figures of Wundt, a longer side of one with a shorter side of the other." Shen adds, "When we conclude that there is a definite tendency toward overestimation according to friendship, we must remember that much the larger part of the errors is due to other unknown factors."

SELF-RATING

Knight and Franzen [2] studied the self-ratings of 110 junior students in a university as compared with their ratings of each other. There was a clear tendency to speak well of themselves in those virtues considered of greater importance by them, and to rate themselves less highly in traits considered less vital. This positive correlation between the relative importance of traits and the amount of each trait a subject rates himself as possessing may well be considered a self-defense mechanism whereby a person tends to think well of himself in what he judges important and underrates himself in less significant items. Common sense shows that all of us would be readier to admit poor memory or poor handwriting

[1] Shen, E. — "The Influence of Friendship upon Personal Ratings," *Journal of Applied Psychology*, 9: 66–68, Mar., 1925.

[2] Knight, F. B., and Franzen, R. H. — "Pitfalls in Rating Schemes," *Journal of Educational Psychology*, 13: 204-213, April, 1922.

than poor judgment or inferior trustworthiness. These writers sum up by saying, "These data amply support us in holding that there is a well-marked tendency for a person to overrate himself when he compares himself with others, and even when the introspective judgment is independent of comparison with others, this tendency still persists." Hollingworth[1] had earlier noted this constant error in self-rating.

Reliability of Ratings

Ratings suffer from rather large unreliability — so large, indeed, that a single rating of one individual by another is next to worthless.

Webb[2] finds the average reliability of a large number of ratings to be .55. By coincidence Voelker[3] reports an average reliability of ratings at .55. Shen[4] has reported 108 reliability coefficients of ratings of various human traits which again average .55. Of course the reliabilities varied from very low to very high, but with reasonable care, reasonable sagacity, and rather close acquaintance, .55 is a good average figure for the ratings of traits of personality. Although Shen's average reliability coefficient was .55, that figure was not the average reliability in separate traits. The average of 13 reliability coefficients in each trait is as follows:

Scholarship	.71
Leadership	.68
Intellectual quickness	.68
Intellectual profoundness	.64
Memory	.55
Resistance	.40
Adaptability	.37
Impulsiveness	.34

[1] Hollingworth, H. L. — *Vocational Psychology*, chs. 6–7.

[2] *Op. cit.*, Tables 4 and 5.

[3] Voelker, P. F. — *The Function of Ideals and Attitudes in Social Education.*

[4] Shen, E. — "The Reliability Coefficient of Personal Ratings," *Journal of Educational Psychology*, 16: 232–236, April, 1925.

H. L. Hollingworth [1] notes the same phenomenon of some traits showing greater objectivity than other traits. Hollingworth's table is herewith presented. The figures are an index of reliability. 100 signifies average reliability. Less than 100 less than average error in rating and over 100 greater than average error in rating.

TABLE 86

THE AMOUNT OF DISAGREEMENT AMONG JUDGES IN ESTIMATING, ON THE BASIS OF ACQUAINTANCE, THE TRAITS OF OTHERS, IN TWO INVESTIGATIONS (FROM HOLLINGWORTH)

TRAIT	RELIABILITY INDEX	
Efficiency	83	
Originality	86	
Perseverance	88	
Quickness	89	A traits
Judgment	89	Close agreement among judges
Clearness	90	
Energy	91	
Will	91	
Mental Balance	96	
Breadth	96	
Leadership	96	
Intensity	99	B traits
Reasonableness	100	Fair agreement among judges
Independence	101	
Refinement	103	
Physical Health	103	
Emotions	105	
Courage	109	
Unselfishness	110	
Integrity	117	C traits
Coöperativeness	119	Poor agreement among judges
Cheerfulness	121	
Kindliness	123	

[1] HOLLINGWORTH, H. L. — *Judging Human Character*, p. 79, 1922. D. Appleton and Company, Publishers.

Of these Hollingworth[1] says,

> The A traits we may designate as "objective," in the sense that they represent reactions to objects and impersonal situations and tasks, and are likely to result in objective products such as inventions, factories, books, bank accounts, salaries, positions, records, etc. These objective products are definite manifestations of the traits in question and they are open to general inspection. The C traits, on the other hand, represent reactions to the presence and character of other persons. They are personal, social, moral; they do not so definitely produce objective products open to general inspection. Instead, they lead mainly to personal and emotional reactions on the part of others; hence we may designate them "subjective traits."

Paterson[2] carries this one step further by suggesting that, insofar as possible, traits be defined objectively. As an example compare these two definitions of leadership suggested by Paterson:

> Subjective definition: Leadership: Rate this executive's initiative, force, self-reliance, decisiveness, tact, ability to inspire men and to command their obedience, loyalty and coöperation.
>
> Objective definition: Leadership: Rate this executive according to the success he has shown in developing a loyal and effective organization by administering justice, inspiring and winning the coöperation of his subordinates.

In connection with reliability Shen[3] has shown that whereas acquaintance with the persons being rated tends to lower the validity of the rating by introducing a constant error, it increases the reliability of the rating by decreasing the variable error. We really know our long-standing friends better although we tend to overrate them. Likewise, although self-rating leads to over- or underrating in desirable or undesirable traits, one is more reliable in rating one's self than in rating others — that is, the variable errors are less. We really know ourselves better than we know any one else.

[1] *Op. cit.*, p. 80 f.

[2] Paterson, D. G. — "Methods of Rating Human Qualities in Psychology and Business," *Annals of American Academy of Political and Social Science*, 110: 81–93, Nov., 1923.

[3] Cf. "The Influence of Friendship upon Personal Ratings."

Because the reliability of ratings is so low, more than one independent judgment should be secured in order to obtain a reliable estimate of any individual. Just how many independent ratings should be obtained may be estimated by means of the Spearman-Brown formula. The accompanying table enables one to know how many independent ratings are needed to obtain certain desired reliabilities. Column 1 is the number of ratings necessary to obtain a reliability of .82 — an average figure for a forty-five-minute standardized test of achievement in high school. Column 2 is the number of ratings necessary to obtain a reliability of .90 — a standard or goal for satisfactory individual diagnosis.

TABLE 87

NUMBER OF RATINGS NECESSARY TO OBTAIN A SPECIFIED RELIABILITY FOR VARIOUS TRAITS

	r_{11}	Number Ratings to Obtain a Reliability of	
		.82	.90
Average ratings	.55	4	8
Scholarship	.71	2	4
Leadership	.68	3	5
Intellectual quickness	.68	3	5
Intellectual profoundness	.64	3	6
Memory	.55	4	8
Persistence	.40	7	14
Adaptability	.37	8	16
Impulsiveness	.34	9	18

It will be well to close this section on rating with Rugg's[1] ringing question and answer:

Can human character be rated on point scales accurately enough for practical purposes in education?

Yes, and — no.

[1] *Op. cit.*

Yes, if the rating is done under conditions as rigorous as the following:

First, if each rating given a person is the average of *three independent* ratings, each one made on a scale as objectively as the man-to-man comparison type of scale.

Second, if the scales on which the ratings are made are comparable and equivalent, having been made in conferences under the instruction of one skilled in rating scale work.

Third, if the three raters are so thoroughly acquainted with the person rated that they are competent to rate.

But these conditions are practically unattainable in public schools. Hence the answer to our original question — no, not by methods so far generally employed, and probably not unless methods of re-rating and checking judgments are carried far beyond present practical possibilities.

Rugg erred on the side of leniency in requiring only three independent ratings. As the above table shows, *eight* is an average number of independent ratings one should obtain if the ratings are to be individually diagnostic.

Objective Tests of Conduct

Psychologists have been experimenting for some time with tests of character, in an endeavor to place the measurement of conduct on the same objective plane as the measurement of ability. The work has met with only a modicum of success, partly because of the inherent difficulties in the problem, which have already been noted. The difficulties may be summarized briefly as (1) testing implies that the person being tested is trying to make a high score, hence the test must be given without the person's being aware that it is a test or what it is a test of; (2) conduct does not usually leave behind a record which may be studied and quantitatively evaluated. Hence if conduct is ever to be measured objectively, it must be a strategic attack. Of the many tests devised one line of development is of special importance. Voelker,[1] wishing to study the influence of ideals and attitudes in social education,

[1] *Op. cit.*

constructed, under the guidance of Thorndike, a battery of tests for the measurement of trustworthiness. Voelker's work was pioneering. He attempted much which has not proved of practical value, yet, like all pioneering, certain of his tests were of considerable promise and paved the way for the work of others. Cady [1] at Stanford University took the most serviceable of the Voelker tests and rearranged them in an endeavor to discover an objective method of estimating juvenile incorrigibility. Besides the Voelker tests, Cady used a modification of a test by Pressey [2] of moral judgment and a revision of a questionnaire originally devised by Woodworth [3] for testing emotional instability or psychopathic tendencies.

Raubenheimer [4] did further work along this line.

RAUBENHEIMER TEST OF POTENTIAL DELINQUENCY

Overstatement, *A.* — This test is a modification of one which has been used by Knight. Fifty book titles were listed on the test blank, twenty of which were fictitious. The score is the number of fictitious titles marked.

The instructions were as follows:

"We want to see who has read the most books. You are to mark a cross on the dotted line in front of every book you have ever read no matter how long ago you read it. After you have finished marking the crosses, count up the number of crosses you have marked. This number will be your score. We want to see who will have the best score. Afterwards we will have you stand up and tell your score. A PERFECT SCORE is 50. Look at the sample."

[1] CADY, V. M. — *The Estimation of Juvenile Incorrigibility,* JOURNAL OF DELINQUENCY MONOGRAPH, NO. 2, April, 1923.

[2] PRESSEY, S. L. and L. C. — "'Cross-out' Tests with Suggestions as to a Group Scale of the Emotions," *Journal of Applied Psychology,* 3 : 138–150, 1919.

[3] WOODWORTH, R. S. See Franz, S. I. — *Handbook for Mental Examination,* ch. 12.

[4] RAUBENHEIMER, A. S. — *An Experimental Study of the Behavior Traits of the Potentially Delinquent Boy,* PSYCHOLOGICAL MONOGRAPHS, NO. 159, 1925.

Sample 1.

. . X . . Anderson's Fairy Tales.

Five of the fifty items follow:

. . . . . . Robinson Crusoe
. . . . . . Little Men
. . . . . . Uncle Remus Stories
. . . . . . The Underground Patrol
. . . . . . White Fang
Etc.

"Now count up your score and put the number you have read in the space here. Remember, a PERFECT SCORE is 50."

MY SCORE IS . . . . . .

Overstatement, *B*. — This test, which is a modification of a test used by Voelker, consists of two parts. The first part calls for a statement of amount of knowledge the child has in regard to 80 items of information. The second part, given later without warning, tests the child's actual information on these same 80 items. The score is the per cent of overstatement or understatement.

Directions: "On the dotted line before each question put *2*, *1*, or *0* to tell how well you know the thing it asks about. Put down *2* if you know it VERY WELL. Put down *1* if you know it FAIRLY WELL. Put down *0* if you KNOW NOTHING ABOUT IT."

First sample: . . . . Can you ride a bicycle?
Second sample: . . . . Can you skate on roller skates?
Third sample: . . . . Can you drive a car?

Sample items follow:

Do you know who discovered America?
Do you know who wrote "Huckleberry Finn"?
Do you know who was the prophet who spent the night in the lions' den?
Etc.

"Now add up your score. A PERFECT SCORE is 160 points."

MY SCORE IS . . . .

"In the second part the subject was told to underline the correct word in such sentences as the following:

America was discovered by Drake Columbus Balboa Cook

"Huckleberry Finn" was written by Alger Dickens Henty Mark Twain

The prophet who spent the night in the lions' den was Daniel Jonah David Joel

Questionable reading preferences. — The child is asked to place a number from *1* to *10* before each of ten book titles to indicate how well he would like to read the book, *1* meaning greatest preference, *10* least. The score is the sum of the squares of the deviations of the individual items from the correct order as determined by competent judges. Two lists were given separately. Part of one of them is as follows:

......A Daring Rescue
......Roy Black, the Master Thief
......Captains of Great Teams
......Hobo Stories
......Running Away with the Circus

Questionable character preferences. — This test consists of eight brief paragraphs, each characterizing a boy. The subject is asked to place before each characterization a number from *1* to *8*, to show how well he would like to have as a chum the boy described. The test is scored in the same manner as test 3. Two such lists of eight items each are given separately. Two of the items in one list are as follows:

Dick joined the Boy Scouts as soon as he was old enough. He did not like it at first; the drill and the rules were hard. Now he is a troop leader and is planning a camp in the mountains next summer.

Ray Stevens is at school now, but he is anxious to get out. He wants to become a taxi-driver. Ray says that taxi-drivers have an easy time; they need not work so hard, and they go about a great deal.

Social attitudes. — Twenty-four things or ideas were named, each followed by four statements expressing various kinds of reactions to the thing or idea presented. The subject is asked to check the one statement that most nearly tells how he feels about it. Score is the number of questionable items checked. Samples are as follows:

CHUMS: ......It is hard to go without them.
......You cannot always trust them.
......They sometimes squeal on you.
......It is best to have them in your gang.

BOY SCOUTS: ......They have too many rules.
......They have to drill too hard; it is no fun.
......They are regular fellows and have lots of fun.
......They are like sissies.

Activity preferences. — Of three things to do, the subject was to check the one thing he liked best. The test contained twelve items of this nature, the score being the number of items in which some other than the "best" response was checked. This test is hardly adapted for use with children of superior social and moral environment. Two of the items are as follows:

A. ......Go camping with Boy Scouts.
......Go around seeing the country, getting lifts as you go.
......Quit school and go with the circus.

B. ......Match pennies and win.
......Have a paper route.
......Win money at the shooting gallery.

Rating the seriousness of offenses. — This is a carefully worked-out version of a test which has long been in use. Ten offenses were listed, covering a wide range in nature and seriousness. They had been selected from a much larger number of offenses actually committed by delinquent boys, and they differed from one another in seriousness by equal steps. The subject was to number them from 1 to 10 according to seriousness. The score was the sum of the squares of the deviations from the correct order. This test was finally omitted from the series used with our gifted children; it correlated too highly with intelligence and some of its items were deemed objectionable on moral grounds. Some of the items were as follows:

(*b*) Sam set fire to the public school which he attended.
(*d*) Bob ran away from home and got a job, getting his room and board from another family.
(*f*) Ted played hookey to go to a circus.
(*j*) Joe entered the house of the people next door and took $2.50.

The reliability coefficients of these seven tests for 13 year olds are given as follows:

1.	Overstatement A, books read	.74
2.	Overstatement B, knowledge claimed	.78
3.	Reading preference	.80
4.	Social preferences	.79
5.	Social attitudes	.77
6.	Activity preferences	.74
7.	Offense ratings	.78

It is seen that in reliability these tests are not far below our standard achievement tests.

Woodworth-Cady Questionnaire

This test was devised by Woodworth in 1918 for use in identifying soldiers having psychotic tendencies. In arranging it the author sifted the leading manuals of psychiatry, made a list of the symptoms most generally agreed upon by competent authorities as being frequently associated with psychopathy, and embodied as many of these symptoms as possible in questions to be answered by yes or no. The war came to an end before the validity of the test could be thoroughly established, but it was given to one regiment and to several groups of normal and psychopathic subjects. With a group of subjects in a private hospital for nervous patients it yielded scores which overlapped only slightly those earned by normal subjects.

In its original form the questionnaire, because of items it contained relating to sex, was not suitable for use with school children. It was later expurgated and revised by Ellen Matthews,[1] who then gave it in the revised form to 1,133 unselected school children and to 376 other children, and made a detailed analysis of the responses to individual items and revised the test. Cady made a number of additional changes based upon the extent to which individual items differentiated children rated by other criteria as superior or inferior in emotional stability. He also arranged it in two forms in order to determine its reliability. In one form the questions were of a positive type throughout; in the second form they were of a negative type. For example, Form *A* contains the question, "Do you always get on well with your teachers or principal?" In Form *B* this item runs, "Do you often have trouble with your teachers or principal?" . . . Two or three items were omitted because they were considered objectionable; for example, the question, "Do you feel that your parents are not really your own?" Certain changes were also made in the "padding"; that is, in the items inserted merely to lull the suspicions of the subject as to the purpose of the test. Finally, small changes were made in the wording of a few of the significant items. . . . The starred items only are scored.

The score is the number of starred items answered correctly.

The "Correct" answers are underlined in the following reproduction of the test.

[1] "A Study of Emotional Stability in Children," *Journal of Delinquency*, 8: 1–40, Jan., 1923.

1.	* Are you happy most of the time?	YES	NO
2.	* Are you afraid in the dark?	YES	NO
3.	* Do you get tired of people easily?	YES	NO
4.	* Are you anxious to get away from school and get a job?	YES	NO
5.	* Have you often fainted away?	YES	NO
6.	Were you ever fond of playing with snakes?	YES	NO
7.	* Did you ever have a nickname you didn't like very well?	YES	NO
8.	* Do you usually enjoy your meals?	YES	NO
9.	* Can you usually sit still without fidgeting?	YES	NO
10.	* Do people say you are disobedient?	YES	NO

Cady[1] claims for this questionnaire a reliability of .75 for a 12–14 age group.

A striking point of this movement in testing character is the renaissance of the questionnaire. The questionnaire promises to do for character measurement what tests have done for ability measurement.

G. Stanley Hall made wide use of the questionnaire in his studies. Hall usually asked adults various questions in an endeavor to obtain *facts* or *information* as to the mental life and development of children. This method of obtaining information is so questionable that it has led Thorndike[2] to comment about the questionnaire method in a way which has generally discredited it as a psychological instrument. The present movement does not use the questionnaire to find out information but to learn more about individuals. Instead of asking questions which require accurate knowledge in order to give an adequate answer, questions are asked concerning the subjects' likes and dislikes, preferences, beliefs, etc. This new type of questionnaire has been found to have satisfactory reliability. One criticism of the new type of questionnaire is that pupils might not give true answers. It is easily conceivable that boys might treat such questionnaires as those suggested by Raubenheimer as a joke and fail to answer them seriously. The answer to that is that they do

[1] *Op. cit.* [2] THORNDIKE, E. L. — *Educational Psychology*, Vol. I, ch. 4.

not. In actual practice these questionnaires are taken in all seriousness, as is shown by the high reliability coefficients. If the questionnaires were treated as a joke two alternative forms of any questionnaire would tend to have zero reliability. The fact is that these questionnaires are treated as seriously as other school work by the large majority of school children.

One should be most careful in referring to questionnaires not to call them tests. They are not tests in the same sense that an arithmetic test is a test. They are used for a different purpose. A test gives a measure of ability. A questionnaire gives a measure of conduct or choice, and where the choice is a moral one, the questionnaire becomes an excellent test of character.

REFERENCES

ALLPORT, F. H. — *Social Psychology.* Houghton Mifflin Company; 453 pp., 1924.

CADY, V. M. — *The Estimation of Juvenile Incorrigibility,* JOURNAL OF DELINQUENCY MONOGRAPH, NO. 2, April, 1923.

CHU, J. P. — *Chinese Students in America: Qualities Associated with their Success,* TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 127, 1922.

FREYD, M. — "The Graphic Rating Scale," *Journal of Educational Psychology,* 14 : 83–102, Feb., 1923.

GALTON, F. — *Inquiries into Human Faculty,* 1883.

HAUGHT, F. B. — "A Scheme for Combining Incomplete Rankings," *Journal of Applied Psychology,* 7 : 168–172, June, 1923.

HOLLINGWORTH, H. L. — *Judging Human Character.* D. Appleton and Company, New York; 268 pp., 1922.

KELLEY, T. L. — *Statistical Methods. Ref.* Ch. 12.

KNIGHT, F. B. — "Analysis of Teaching and Teachers," *Journal of Educational Research,* 10 : 224, Oct., 1924.

— "The Effect of the 'Acquaintance Factor' upon Personal Judgments," *Journal of Educational Psychology,* 14 : 129–142, March, 1923.

KNIGHT, F. B., and FRANZEN, R. H. — "Pitfalls in Rating Schemes," *Journal of Educational Psychology,* 13 : 204–213, April, 1922.

MATTHEWS, E. — "A Study of Emotional Stability in Children," *Journal of Delinquency,* 8 : 1–40, Jan., 1923.

MINER, J. B. — "The Evaluation of a Method for Finely Graduated Estimates of Abilities," *Journal of Applied Psychology*, 1 : 123–133, June, 1917.

NORSWORTHY, N. — "The Validity of Judgments of Character" in *Essays in Honor of William James*. Second edition, pp. 542–552, 1910.

PATERSON, D. G. — "Methods of Rating Human Qualities in Psychology and Business," *Annals American Academy of Political and Social Science*, 110 : 81–93, Nov., 1923.

PERSONNEL SYSTEM — U. S. Army Rating Scale C. C. P. 400. 1 : 559–581, 1919.

RAUBENHEIMER, A. S. — *An Experimental Study of Some Behavior Traits of the Potentially Delinquent Boy*, PSYCHOLOGICAL MONOGRAPHS, NO. 159, 1925.

RUGG, H. O. — "Is the Rating of Human Character Practicable?" *Journal of Educational Psychology*, 12 : 425–438, 485–501, 13 : 30–42, 81–93, Nov., Dec., 1921; Jan., Feb., 1922.

SHEN, E. — "The Influence of Friendship upon Personal Ratings," *Journal of Applied Psychology*, 9 : 66–68, March, 1925.

— "The Reliability Coefficient of Personal Ratings," *Journal of Educational Psychology*, 16 : 232–236, April, 1925.

— "The Validity of Self-Estimate," *Journal of Educational Psychology*, 16 : 104–107, Feb., 1925.

SYMONDS, P. M. — "Notes on Rating," *Journal of Applied Psychology*, 9 : 188–195, June, 1925.

— "On the Loss of Reliability in Ratings Due to the Coarseness of the Scale," *Journal of Experimental Psychology*, 7 : 456–461, Dec., 1924.

TERMAN, L. M., and others — *Genetic Studies of Genius*, Vol. I, ch. xvii. Stanford University, Stanford; 648 pp., 1925.

THORNDIKE, E. L. — "A Constant Error in Psychological Rating," *Journal of Applied Psychology*, 4 : 25–29, March, 1920.

THORNDIKE, E. L., and others — *Psychology of Algebra*. *Ref.* Ch. 3.

VOELKER, P. F. — *The Function of Ideals and Attitudes in Social Education; An Experimental Study*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 112. Teachers College, Columbia University, New York, 1921.

WEBB, E. — *Character and Intelligence*, BRITISH JOURNAL OF PSYCHOLOGY MONOGRAPHS, NO. 3, 1915.

WOODWORTH, R. S. — See Franz, S. I.: *Handbook for Mental Examination*, ch. 12.

CHAPTER XVII

PROGNOSIS TESTS

Prognostic testing is the chapter of romance in the testing movement. Science holds prediction as its most important aim and prognosis is the ultimate aim of endeavor in the scientific study of education.

The theory underlying the construction of a prognostic test is quite simple. In its present state it is almost a trial-and-error matter. To construct a prognostic test of ability to learn demonstrative geometry, one would give a battery of tests to a group of pupils who are about to start their study of geometry. These tests would have to be selected almost at random. He would try out the standardized intelligence tests, being careful to sample adequately all forms of intelligence tests such as completion, analogies, etc.; and all kinds of material such as verbal, numerical, geometric forms, etc. He would test knowledge of geometrical concepts, forms and facts, and ability to compute areas and volumes, and to do simple constructions, — all of which matters have been learned in previous schooling or in life's experiences. By means of easy exercises he would also test ability to grasp quickly the principles of a geometric proof. After giving and scoring such a miscellaneous set of tests nothing further would be done until the pupils had studied demonstrative geometry for a year, then a thorough test would be made of geometry achievement. The correlation would be found between each of the tests given before the study of geometry was commenced and the achievement in geometry after a year of study. Those tests which correlate

most highly with achievement are considered to have the most predictive value and may be assembled to form the prognosis test. That, in brief, is the principle of the construction of a prognosis test. It is a test which actually is made to predict as well as possible by trial. The check on its prognostic qualities is found in the coefficients of correlation.

The idea of prognostic testing originated with T. L. Kelley, who was first to construct tests to predict achievement. This work is reported in Kelley's *Educational Guidance* published in 1914. Kelley stated his aim as follows: "The endeavor of this study is to predict with a known, and as high as possible, degree of accuracy the capacity of the pupil to carry a prospective high school course." As worked out the problem is not actually as simple as indicated above in the case of geometry. Not only does one wish to know the correlation of the separate tests with expected achievement, but also the correlation of each test with each other. It might come about that two tests which correlate highly with expected achievement also correlate highly with each other, that is, they might measure almost the same function. If so, then in combination they would predict later achievement no better than either one taken singly. That is, to select a battery of tests that will make the best prognosis of achievement, one should select tests that not only correlate highly with expected achievement but slightly with one another. To make this selection of tests one uses partial correlations. The importance of the tests for the prognosis of prediction is shown by the regression equation, an equation of the form

$$X_m = c_o + c_1 X_1 + c_2 X_2 + c_3 X_3 + \text{etc.}$$

where c_o, c_1, c_2 are coefficients showing the weights the tests should have in order to make the best possible prediction.

Kelley[1] devised four tests which he hoped would be predictive

[1] *Op. cit.*, Section V.

of ability in first-year high school. One test he calls an algebra test (called an algebra test simply because it was given to classes just starting in algebra), although it is in reality a test of ability in arithmetic. Eight of the problems test the thoroughness of the preceding preparation of the pupil in arithmetic, three test ability to understand the written statement of a problem and to deal with negative magnitudes; and three demand a process of thinking which is very closely related to simple algebraic reasoning. Another test he calls a geometry test to be given to beginning geometry classes. One problem is to test ability to image geometrical forms and movement; five problems are locus problems, demanding a common sense interpretation of geometric facts; one problem is to test the pupil's previous preparation in the line of ratio and proportion; three problems are designed to test the pupil's ability to handle *reductio ad absurdum* proofs and converse propositions. The English test consisted of a poem read aloud to the class followed by these questions to be answered by the pupils: (1) State one or two things about the story which you particularly liked; (2) Write an account of the story as fully as you can remember it. These answers were graded according to the following points: (*a*) Valuation of the essential ideas of the poem, (*b*) Accuracy and extent of description, (*c*) Written expression, (*d*) Dramatization. The history test was similar. A passage of history text was read to the class, after which the pupils answered these questions: (1) What do you think was probably the next most important event, (2) Describe the character and appearance of Garibaldi, (3) Beginning with the departure of Garibaldi and his men from Genoa write a detailed account of as much of the story as you have time for. Each paper was scored according to the same points as the English test.

There are two important facts to note from Kelley's results. First, although algebra is best predicted by the algebra test, the English and history tests are not entirely worthless in predicting

algebraic ability. As a matter of fact, achievement in history is better predicted by the mathematics and English tests alone than by the history tests alone. This fact demonstrates that the various school subjects are interrelated and that one may often predict success in one subject by a test in another subject called by an entirely different name. Probably the unitary abilities making for success in school subjects cut across the subjects without respect to material or form.

Second, the size of the correlations themselves are of significance. The algebra test correlates with achievement in algebra only .47; the geometry test with achievement in geometry .42; the English test with achievement in English .44; and the history test with achievement in history only .31. In fact the multiple correlations resulting from the best combination of tests is only .48, .43, and .46. For the purposes of prognosis these may be considered low correlations. Kelley's tests, although ingeniously constructed, probably do not possess the reliability that more recent objective tests possess.

I am not aware that Kelley's tests have been subsequently used or that any practical use has been made of his results. His contribution is in setting up the problem and developing the technique for its solution.

A second attempt in the construction of a prognostic test was that of Dr. Agnes Rogers,[1] beginning in 1916 and published in 1918.[2] Her study was an attempt to analyze mathematical ability, and the development of a battery of prognostic tests was the major outcome of her work. The tests are not strictly prognostic in the sense that they can be used on pupils who have had no instruction in algebra or geometry. The preliminary testing was

[1] *Experimental Tests of Mathematical Ability and Their Prognostic Value*, pp. 72, 74.

[2] Published by Bureau of Publications, Teachers College, Columbia University, as "Rogers Test of Mathematical Ability."

made on two groups, one having had five months of algebra and the other five months of algebra and intuitive geometry. The tests contained material which no one could do successfully without some training in algebra and geometry. In selecting the tests to use in the preliminary testing Rogers endeavored to touch as many forms of mathematical achievement as possible. Six were tests of algebraic ability, five were tests of geometrical ability. A third group consisted of tests of language ability. The criterion in the case of the Rogers tests was not a measure of achievement in mathematics, but a composite of all of the mathematical tests which were used, weighted empirically according to the importance of the ability measured, the reliability of the test, and the standard deviation of the test scores. No partial or multiple coefficients of correlation were computed, but the correlations were corrected for inaccuracies in the measures (correction for attenuation). The tests[1] which correlated most highly with the criterion were

TABLE 88

CORRELATIONS OF TESTS WITH CRITERION OF MATHEMATICAL ABILITY (FROM ROGERS)

Test	Crude Coefficients	Test	Corrected Coefficients
Algebraic computation .	.69	Algebraic computation	.81
Interpolation	.66	Missing steps in series	.78
Missing steps in series .	.63	Geometry	.76
Geometry	.63	Matching equations and problems . .	.76
Superposition	.57	Interpolation . . .	.75
Matching equations and problems . . .	.57	Arithmetic problems .	.74

These tests were not used, however, to form the prognostic battery.

[1] ROGERS, A. — *Op. cit.*, p. 113.

This battery of six tests, known as the "Rogers Sextet," consists of the following tests:

Algebraic Computation
Interpolation
Geometry
Superposition (a modified form of the Thurstone Spatial Relations Test)
Mixed Relations
Trabue Language Scale L and J.

Rogers states that they are "designed to measure the more important phases of mathematical capacity demanded by high school mathematics and in particular the ability to manipulate numerical and algebraic symbols, the ability to grasp and handle spatial relations, and the ability to deal effectively with words."

The correlation of the sextet with school marks is as follows:[1]

TABLE 89

CORRELATION OF ROGERS TEST OF MATHEMATICAL ABILITY WITH SCHOOL MARKS

	CRUDE COEFFICIENTS	CORRECTED COEFFICIENTS
School 1	.62	.79
School 2	.82	.92

These correlations are considerably higher than those found by Kelley. This is in part due to the greater refinement and consequent greater reliability of the tests.

As prognostic tests the Rogers sextet is to-day quite inadequate. The following criticisms may be mentioned: (1) In an attempt to measure mathematical ability the Rogers sextet of tests straddled many different kinds of abilities. Algebra ability is no more like

[1] ROGERS, A. — *Op. cit.*, p. 93.

geometry ability than either is like English ability, as witnessed by these correlations found by Rogers.[1]

TABLE 90

INTERCORRELATION OF ALGEBRAIC ABILITY, GEOMETRY ABILITY AND VERBAL ABILITY (FROM ROGERS)

	GEOMETRY ABILITY	VERBAL ABILITY
Algebraic Ability . .	.54	.57
Geometry Ability . .		.59

Before Rogers' work there was a current superstition that mathematical ability was unique and specialized. As the result of her work, we know that one can estimate capacity for mathematical achievement nearly as well by a general test of intelligence as by any special set of six tests. There is very little reason for attempting to construct a prognostic test of general mathematical ability. Perhaps "algebraic ability" and "geometric ability" may be slightly more specialized than mathematical ability. (2) The tests are not true prognostic tests in the sense that they can be given before the study of algebra or geometry is commenced. Two of the tests definitely depend on preliminary training in algebra and geometry. A true prognostic test presupposes no training whatever in the ability to be tested.

There is only one form of the test, but each test in the booklet is in duplicate. Several of the tests have practice tests, which tends to increase the reliability.

When the tests are administered to those with more than five months of algebra or who have had demonstrative geometry, it is recommended that the time limit on the algebraic computation and the geometry tests be reduced. Complete instructions for giving are included in the manual.

[1] *Op. cit.*, p. 80.

The time allowance is as follows:

TABLE 91

ROGERS TEST OF MATHEMATICAL ABILITY: TESTS AND TIME ALLOWANCE

	TIME FOR EXPLANATION IN MINUTES	TIME FOR TEST IN MINUTES
Geometry test . .	8	22
Algebraic computation test .	¼	1 } 3 2 } 7
Interpolation test . . .	5	1 } 8 2 } 5
Superposition test	5	1 } 1 2 } 1
Trabue Language Scales	2	L } 5 J } 5
Mixed relations test	3	3
	23¼	60

Part of the chaos resulting from using a prognostic test that cannot be administered before the pupil commences the subject is evidenced in the norms. Four sets of norms are given corresponding to classes which have taken the test at various stages of progress in mathematics. Naturally this is somewhat confusing. A prognostic test should have standards. Rogers has not provided standards, except roughly, even with the norms which she presents.

A study has been made of prognostic tests in Latin by Allen.[1] This is probably the first study to use the full technique for the construction of a prognostic battery. Twenty-one psychological tests were given to 443 boys at the opening of school in 1921. This group studied Latin during a semester. Three hundred sixty-four boys remained in the group to be tested in January, 1922, by an extensive battery of tests designed by T. H. Briggs

[1] ALLEN, W. S. — *A Study in Latin Prognosis*, p. 13.

to measure Latin achievement. Pupils' class marks were also used as a secondary criterion. The tests which Allen found most significant with the correlations with the criterion are:

TABLE 92

CORRELATION OF TESTS WITH CRITERION OF LATIN ABILITY (FROM ALLEN)

		Latin Test Criterion
A	Briggs' Analogies *A*	.526
	Thorndike Test of Word Knowledge *A*	.369
	Rogers Interpolation Evens	.355
B	Briggs' Analogies *B*	.559
	Thorndike Test of Word Knowledge *B*	.377
	Rogers Interpolation Odds	.398
C	Briggs' Analogies Combined	.564
	Thorndike Test of Word Knowledge Combined	.389
	Rogers Interpolation Combined	.398

These three groups combined may be known as prognosis tests *A*, *B*, and *C* and they have correlations as follows:

TABLE 93

CORRELATIONS OF ALLEN LATIN PROGNOSIS TESTS

	Latin Test Criterion	Latin Test Criterion	Latin Marks	English Marks	Math. Marks
	1st Semester	2d Semester	1st Semester		
Prognosis Test A	.559	.482			
Prognosis Test B	.578	.458			
Prognosis Test C	.588	.476	.481	.548	.533

It is interesting to note that Briggs' Analogies correlates alone very nearly as highly with the test criterion as the complete battery. Allen gives the following list of prognosis tests with times and correlations.

TABLE 94

ALLEN LATIN PROGNOSIS TEST: TESTS, TIME ALLOWANCE, AND CORRELATION WITH CRITERION

	TIME	CORRELATION WITH CRITERION
Briggs' Analogies Beta	13 min.	.559
Briggs' Analogies Alpha and Beta	26 min.	.564
Prognosis Test A	35 min.	.559
Prognosis Test B	35 min.	.578
Combined Prognosis Test . .	70 min.	.588

In choosing which test or combination of tests to use one needs to figure whether the added correlation gained is worth the added time spent.

Clem,[1] using Allen's tests and the same criterion, found the correlation of Allen's combined prognosis test with the criterion to be .563 for 80 girls and .667 for 103 boys.

Allen has not published his tests as a battery and consequently they have not been used since as a battery for Latin prognosis. But his work was very thorough and sound, and he discovered an excellent battery of tests for the purpose of predicting Latin achievement. He also paved the way for future workers.

There are other tests that are on the market under the name of prognosis or aptitude tests.[2] Few of these tests have employed the technique of correlation in their construction, and hence can-

[1] CLEM, O. M. — *Detailed Factors in Latin Prognosis*, p. 42.

[2] ORLEANS, J. S., and SOLOMON, M., have recently published through the World Book Company a Latin Prognosis Test. This test is of the quick learning type. It consists of a number of brief Latin lessons involving matters of vocabulary and syntax. Correlations as high as .80 with a criterion of Latin ability have been obtained.

not be said to have the validity of the tests already described. Most of these tests were assembled after analysis of the activities or subjects to be predicted. R. E. Hoke has assembled a prognostic test of stenographic ability. The basis for the selection of the seven elements in the test was an attempted analysis of the work of the stenographer into its component mental and motor functions. Following this analysis the effort was made to provide an appropriate method of measuring each of the capacities needed in stenography. The tests are (1) motor reaction, (2) speed of writing, (3) quality of writing, (4) speed of reading, (5) memory, (6) spelling, (7) symbols. The value of this group of tests for prognostic purposes should be held in abeyance pending their actual trial as predictive instruments.

Wilkins has constructed a prognosis test in modern languages which consists of a seeing and writing test; a hearing and writing test; a memory test; and a test of grammatical concepts. For individual testing a seeing-speaking test and a hearing-speaking test has been added. Wilkins has produced no evidence as to the value of his tests and until objective evidence is at hand judgment concerning them must be subject to the whim of individual fancy.

Prognostic tests have considerable promise. Teachers would welcome such tests if they were available and valid. They could be used as a basis for sectioning into homogeneous groups which would be likely to make great or small progress in a subject during a semester or year. They would also find a place in the guidance functions of the school in helping pupils decide whether or not to pursue a given subject. The construction of prognostic tests is still in its infancy, but perhaps we have moved far enough that a few generalizations may be hazarded. In the first place, prognostic tests have shown us that the different subjects in school are not specialized to nearly the degree once supposed. Allen [1] found,

[1] Allen, W. S. — *A Study in Latin Prognosis.*

for instance, that the very tests which he assembled to predict Latin ability predicted marks in English and mathematics better than marks in Latin. A good intelligence test is a good prognostic test in almost any subject. Secondly, it is highly probable that those tests are best for prognostic purposes which have some community of subject matter with the subject to be predicted. Ability in algebra may be predicted by means of tests in arithmetic problems and number series completion; language ability may be predicted by the Briggs' analogies test. It will be noted that these tests are also good intelligence tests. Thirdly, tests that have been assembled because of resemblance of their tasks to the activities of the subject to be predicted should be questioned until the tests themselves pass the test of usefulness. The names by which we label tests often have little relationship to the abilities actually being tested. Courtis and Thorndike,[1] for instance, found that a certain "arithmetic test" measured to large extent speed of writing. Superficial analogy cannot take the place of or be substituted for correlation analysis.

Gates and La Salle[2] have made a significant study of the predictive value of tests in the elementary school, and some of their conclusions may be appropriated and applied to prognosis on higher levels. The "National Intelligence Test," the Stanford Revision of the Binet, and several subject-matter tests were given in the Scarborough School, and the subject-matter tests were repeated at intervals of 4, 8, 12, 16, and 20 months. Their findings are:

1. The obtained correlations between the "National Intelligence Tests" and achievement tests, and the intercorrelations

[1] Courtis, S. A., and Thorndike, E. L. — "Correction Formulæ for Addition Tests," *Teachers College Record*, 21: 1–24, Jan., 1920.

[2] Gates, A. I., and La Salle, J. — "The Relative Predictive Values of Certain Intelligence and Educational Tests Together with a Study of Effect of Educational Achievement upon Intelligence Test Scores," *Journal of Educational Psychology*, 14: 517–539, Dec., 1923.

of achievement tests decrease steadily as the intervals increase from 0 to 20 months.

TABLE 95

CORRELATIONS OF NATIONAL INTELLIGENCE TEST WITH AVERAGE OF FOUR SCHOLASTIC TESTS

INTERVAL OF MONTHS	0	4	8	12	16	20
Correlations . .	.868	.828	.838	.848	.790	.755

2. The Stanford-Binet predicts achievement 20 months later about as well as for shorter periods.

TABLE 96

CORRELATIONS OF STANFORD-BINET WITH AVERAGE OF FOUR SCHOLASTIC TESTS

INTERVAL OF MONTHS	0	4	8	12	16	20
Correlations . .	.69	.723	.725	.718	.703	.713

3. When corrections for the unreliability of tests are made, it appears that the best means of predicting achievement in a particular subject, — reading, spelling, or arithmetic, — is to use a test of that subject itself.

4. The usefulness of an intelligence test, because of the universality of its currency, is indicated, however.

Probably very much more could be done with existing achievement tests than is now done in making predictions of future achievement in the high school.

REFERENCES

ALLEN, W. S. — *A Study in Latin Prognosis.* TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 135. Teachers College, Columbia University, New York; 41 pp., 1923.

CLEM, O. M. — *Detailed Factors in Latin Prognosis.* TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 144. Teachers College, Columbia University, New York; 60 pp., 1924.

— "Latin Prognosis: A Study of the Detailed Factors of Individual Pupils," *Journal of Educational Psychology*, 16 : 160–169, March, 1925.

FULLER, F. D. — "Study of the Hoke Prognostic Test of Stenography," *Educational Research Bulletin* of the Los Angeles City Schools, 5 : 5–6, March, 1926.

GATES, A. I., and LA SALLE, J. — "The Relative Predictive Values of Certain Intelligence and Educational Tests Together with a Study of the Effect of Educational Achievement upon Intelligence Test Scores," *Journal of Educational Psychology*, 14 : 517–539, Dec., 1923.

HENMON, V. A. C. — "Prognosis of Ability in School Subjects," *Educational Research Bulletin* of the Los Angeles City Schools, 5 : 23, April, 1926.

HULL, C. L., and LIMP, C. E. — "The Differentiation of the Aptitudes of an Individual by Means of Test Batteries," *Journal of Educational Psychology*, 16 : 73–88, Feb., 1925.

LIMP, C. E. — "The Use of the Regression Equation in Determining the Aptitudes of an Individual," *Journal of Educational Psychology*, 16 : 414–418, Sept., 1925.

KELLEY, T. L. — *Educational Guidance: An Experimental Study in the Analysis and Prediction of High School Pupils.* TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 71. New York, 1914.

OTIS, A. S. — *Statistical Method in Educational Measurement. Ref.* Ch. 12.

ROGERS, A. L. — *Experimental Tests of Mathematical Ability and Their Prognostic Value.* TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION NO. 89. Teachers College, Columbia University, New York; 120 pp., 1918.

CHAPTER XVIII

TESTS OF CLERICAL AND MECHANICAL ABILITY

Clerical and mechanical tests are full of promise as tools to use in educational guidance, but the tests themselves are most disappointing, notwithstanding the fact that much work has been done upon them. Only too often tests have been assembled and named without being tried out by persons trained in statistical methods. The emphasis throughout this book is that before a test can be recommended one must know what it does from trial. This is particularly true with clerical and mechanical tests.

The tests to be described are known as "aptitude tests," which may be considered merely another name for prognostic tests. Concerning aptitude tests Thurstone says,

> An aptitude test is that particular kind of psychological test which is intended to measure the ability to learn some special thing. An aptitude test for telegraphers is a test which is intended to measure a person's ability to learn telegraphy. The aptitude test for telegraphers is not, of course, a test involving telegraphy, because the competitor is not supposed to know anything about telegraphy when he takes an aptitude test. . . .

Clerical Tests

THE THURSTONE CLERICAL TEST

The Thurstone Clerical Test was the first clerical aptitude test to be published.[1] It was first reported in 1919. It consists of eight parts, as follows:

[1] Published by the World Book Company in 1920.

I. Checking the errors in addition and subtraction.
II. Underscoring misspelled words in a passage taken from Arnold Bennett's "Mental Efficiency."
III. Canceling the four letters X, Y, U, and C.
IV. A code-learning substitution test.
V. A test in alphabetizing.
VI. A test in classifying insurance policies in three classes according to kind of policy, date of issue, and the amount.
VII. An arithmetic test.
VIII. Matching proverbs.

This is an omnibus test with a maximum time of 90 minutes, and an average time of taking 40 minutes.

The scoring is in terms of both speed and accuracy. Speed is the number of minutes spent in taking the test — accuracy is the number of correct items. The test was tried out in many large industrial firms. Using as a criterion employees in an invoice company rated in five classes according to the grade of office work in which they were employed, Thurstone found a correlation of .50 between this measure of efficiency and accuracy in the test; .42 between efficiency and speed and .61 between efficiency and speed and accuracy combined. Concerning this test Thurstone says that it endeavors to fit the appropriate intelligence level of clerical workers and to appeal to the applicant for an office position. He adds, "I do not believe that office work has any special abilities that have so far been demonstrated." [1]

THE I. E. R. GENERAL CLERICAL TESTS

Thorndike in 1914 was experimenting with clerical tests, but the war broke into the continuity of a long-time experiment which he had planned and the test was not published. The work was taken

[1] THURSTONE, L. L. — "Standard Tests for Office Clerks," *Journal of Applied Psychology*, 3: 248–251, Sept., 1919.

up by the vocational guidance inquiry of the Institute of Educational Research in 1921 by Toops. Both the Thorndike test and a clerical test devised by Toops at Camp Grant were tried out. The one by Toops was developed into what is known as the "I. E. R. General Clerical Test C–1." This test correlates rather well with intelligence. It has a reliability of .82 when given in a commercial high school and correlates about .70 with success in business school work in stenography, typing, and bookkeeping. Concerning this latter, Toops[1] says that the correlations may be "due to its value as a test of ability to deal with ideas rather than to its value as a test of ability to deal with clerical items and procedures." However, Toops was not able to get correlations over .15 between this test and ratings on nearly 300 clerks in a large business establishment. Perhaps the criterion ratings are to be blamed for these low correlations.

The Thorndike tests were developed into the "I. E. R. Routine Clerical Test C–2." This battery is made up of six tests which were found by Thorndike to correlate with clerical ability, four of which were found by McCall to correlate very slightly with general intelligence in a grade population. Test C–2 correlates less closely with an abstract verbal test than test C–1, also less closely with school success. But C–2 showed a correlation of .38 with clerical work, whereas C–1 showed a correlation of .40 (N = 73). Concerning these results Toops[2] says:

> Clerical ability, so far as can be determined from our investigation, is not unique in the sense that mechanical ability is, that is, clerical ability is highly related to intelligence, although there evidently are some specific factors making for clerical success that are not measured by intelligence. It has been found that there is a hierarchy in the levels of clerical ability.

[1] Toops, H. A. — *Tests for Vocational Guidance of Children Thirteen to Sixteen*, p. 96.

[2] Toops, H. A. — "The Technique of Vocational Guidance as Investigated by the Institute of Educational Research," *The National Vocational Guidance Bulletin, no. 6*, pp. 86–93, Jan., 1923.

Routine clerical workers make considerably lower scores than accountants and minor executives. Only a small percentage of New York City public school boys and girls make test scores as high as the average business college student. Thus at the age of 13 or 14, clerical ability is not sharply distinguished from general academic or intellectual ability. It might be added that " I. E. R. General Clerical Test C–1 " is the better test to use with high-grade clerical positions, whereas " I. E. R. Routine Clerical Test C–2 " is the better test to use with the more routine work.

RUGGLES TEST OF CLERICAL ABILITY

Ruggles,[1] while working with the Wisconsin Civil Service Commission, assembled a battery of six tests for predicting clerical ability. These tests were the result of a thoroughgoing analysis of the skills required in clerical occupations.

A careful study of clerical work, especially as illustrated by the clerical performance scale cards, resulted in the selection of six fundamental operations, with varying possible ranges of difficulty within themselves. Those determined upon as practically covering the entire cross-section of clerical work are as follows:

1. Checking of material for correspondence with copy or similar material, with which it is compared for sameness.
2. Selection of data on the basis of comparison.
3. Locating in an alphabetical index.
4. Classification or sorting.
5. Tabulation or selection and transfer of data.
6. Mathematical computations.

Ruggles [2] also attempted by correlation methods to determine the validity of his tests. Eight groups of clerks were used. Ratings were given by supervisors as overseers on both *aptitude* and *efficiency*. Ruggles does not give correlations of the battery of tests as a whole with his criteria, but gives the correlations of each separate test in great detail. These average correlations range from .33 for test 4 to .46 for test 6. These same tests correlate with "Army Alpha" from .15 to .51 although the " Army Alpha "

[1] RUGGLES, A. W. — *A Diagnostic Test of Aptitude for Clerical Office Work*, p. 15.
[2] *Op. cit.*, p. 38.

correlations are low in each case except for test 4. Ruggles' thorough analysis of each of his tests is an excellent basis for further work for workers with clerical tests.

BUREAU OF PUBLIC PERSONNEL ADMINISTRATION TESTS

The Bureau of Public Personnel Administration is now doing a unique work and is making a valuable contribution in furthering the work of making and validating tests in industry. One of the first pieces of research undertaken by this bureau was a revision of civil service examinations of clerical workers to embody in a new test modern principles of test construction. This work was undertaken by O'Rourke. Thurstone [1] followed with an extensive analysis of existing clerical tests which reviewed the work that had been done up to that time (1923) and paved the way to further refinement of these tests. The latest revision is that sponsored by the Bureau of Public Personnel Administration. In the article describing the test no one is made directly responsible for its construction, although credit is given to Stenquist, Short, Rease, Messich, Moran, Alegeyer, Thorndike, Woodworth, and Toops for suggestions and criticisms. It is entitled "Partially Standardized Tests for Junior Clerks." This test attempts "a better balance between tests of abstract intelligence and tests of clerical aptitudes and ability which can be obtained by separating these two kinds of material, by using one of the standardized tests of abstract intelligence, and by giving each of the tests or groups of tests the relative weights which it seems to deserve."

An outline of the test follows:

Test 1. Copying and adding.
(30 items are to be selected from a much larger list of items and copied into five columns; the sum of each column is then found.)

[1] THURSTONE, L. L. — "A Comparative Study of Clerical Facts," *Public Personnel Studies*, 1, pp. 2–6, 1923.

Test 2. Business information.

(Section A consists of 20 questions in the multiple choice form, each with four answers from which the competitor chooses the best one. Section B consists of 25 statements in the completion form each of which requires the competitor to write in a word or phrase to supply missing information. Each question or statement relates to business or office transactions, mostly of a type familiar to the adult without clerical experience and about which the person with clerical aptitude is apt to have some information.)

Test 3. Alphabetical filing.

(This test calls upon the competitor to insert 20 names in the proper places in a list of 72 names in accordance with rules as to alphabetical order which are given.)

Test 4. Understanding of printed literature dealing with office problems.

(This test consists of 10 questions relating to content of a quoted section dealing with business administration or office management.)

Test 5. Social intelligence.

(This test consists of 20 questions in the multiple choice form, each with four answers from which the competitor chooses the best one. Each question deals with relations with fellow employees or the public.)

Test 6. Understanding and following written directions.

(The competitor is asked to carry out 10 written directions relative to the checking of certain items in a table.)

Test 7. Completion of tabulations.

(In each of two tables the competitor is asked to insert 5 missing items requiring both reasoning and calculations; each table contains 8 columns and there are 5 items in each column.)

Test 8. Memory for oral directions.

(A set of directions such as might be given to a clerk are read and the competitor is asked to answer in writing 5 questions as to their meaning. In addition, a list of 15 items is read and the competitor is asked to record in writing as many as he can remember up to 10.)

Test 9. Abstract intelligence.

(One of the five alternative forms of the Army Alpha Intelligence

Test is given; a minimum score of 75 on the Army Alpha scale is required.)

Test 10. Education and experience.[1]

Tests 1–8 have been printed in a booklet.

We are evidently in the midst of a movement to improve clerical aptitude tests, and this Public Personnel test represents merely a step in the movement. It is described in such detail, however, because it is the latest revision embodying all of the previous experience up to the present time.

Mechanical Aptitude Tests

STENQUIST ASSEMBLY TEST OF MECHANICAL ABILITY

The test which has been given the most trial as a measure of mechanical aptitude is the "Stenquist Assembly Test of Mechanical Ability,"[2] devised by J. L. Stenquist. Stenquist was working on this test as early as 1915. It was used extensively on illiterates in the army during the war. The test consists of ten common mechanical contrivances which have been taken apart and which the testee must assemble. There are two series, I and II, not exactly equal in difficulty although results on the two series may be made comparable by the use of extensive norms provided by Stenquist.

The contents of the two series are as follows:

Series I

1.	Cupboard catch	6.	Shut-off
2.	Clothes pin	7.	Wire stopper
3.	Hunt paper clip	8.	Push button
4.	Chain	9.	Lock No. 1
5.	Bicycle bell	10.	Trap

[1] No author given — "Partially Standardized Tests for Junior Clerks," *Public Personnel Studies*, 3: 346–371, 1925. By permission of the Bureau of Public Personnel Administration.

[2] Made and sold by C. H. Stoelting and Company.

Series II

1. Pistol
2. Elbow catch
3. Rope coupling
4. Expansion nut
5. Sash fastener
6. Expansion rubber stopper
7. Calipers
8. Defiance paper clip
9. Double action hinge
10. Lock No. 2

These now come in a convenient wooden box, each object partitioned off from its neighbor by a wooden piece. These boxes may be strapped together in series of eight and carried about like a suitcase. The contrivances were selected as being "fool proof," easy to score, easy to assemble, and procurable in the general trade. In earlier forms of the test the pupil was expected to copy a model, but in the final revision there is no model to follow, — the task is one of assembling without a model.

The Stenquist Assembling tests may be given in groups. As many can be tested at one time as there are boxes for. In giving the test the score sheets are to be first distributed and filled out by each pupil. Then the boxes are distributed but not opened until the directions have been read. The directions are printed on the cover of each box and are also read aloud by the examiner. Thirty minutes are allowed for the test. Each pupil should close his box with his record sheet inside when he is finished. If a pupil finishes early, a bonus is added to his score. The time of finishing (if before 30 minutes) should be noted on the pupil's scoring sheet. One half point for each minute remaining up to 30, may be added to the score.

The examiner acts as scorer immediately after the test. Sitting at his desk the scorer opens one box, examines model A and records score on Record Sheet. Then he examines model B and so on until the box is scored. Boys may act as assistants to disassemble the contrivances. Skilful boys can be taught to do this quickly and accurately, although the examiner should exert constant supervision.

A score of *10* is given to each complete and correctly assembled contrivance. Partial scores may be given for each contrivance. A convenient scoring sheet for the "Stenquist Mechanical Assembly Test" has been prepared and is for sale by the Institute of Educational Research of Teachers College. This scoring sheet gives standards for determining the partial scores for each contrivance.

Stenquist provides tables for converting the scores into T-scores. Norms for the assembly test are given in the following table:

TABLE 97

STENQUIST MECHANICAL ASSEMBLY TEST: AGE NORMS

AGE	SERIES I	SERIES II
12	30.1	28
13	37	42
14	41	43
15	49	43

More important than these age norms are the percentile norms which Stenquist gives in his monograph.

Mechanical ability seems to be quite distinct from abstract-verbal ability. Both Stenquist and Toops agree that mechanical ability has little relationship to general intelligence (as measured by our standardized intelligence tests). In equal age groups the correlation between mechanical ability and general intelligence is only about .20. Toops[1] says, "This means that a large proportion of pupils who fail in the elementary school may yet be able to make fair or good success in trade work. This conclusion is of immense value for educational and vocational guidance." The reliability of the assembly test is about .7 for grade or high school groups. Stenquist finds that the Assembly test correlates

[1] TOOPS, H. A. — *The Technique of Vocational Guidance.*

with shop teacher rank as follows: .83, .80, .42, .81, .90, .88. These facts may be briefly summarized as follows: The "Stenquist Assembly Test" has substantial reliability, correlates low with general intelligence and high with work in trade courses. These results have led Stenquist [1] to question our right to define general intelligence by our abstract paper-and-pencil tests. His claim is that the boy with high mechanical ability but low paper-and-pencil I Q has just as much right to claim "high intelligence" as the boy who has low mechanical ability but high paper-and-pencil I Q.

Toops [2] gives as his opinion, after a thorough study of various types of tests of mechanical ability — "We are reasonably certain that the Stenquist Assembly Test is to date the most important single test contribution to the measurement of general mechanical ability."

STENQUIST MECHANICAL APTITUDE TESTS

Stenquist has also constructed a paper-and-pencil mechanical ability test which is published under the title "Stenquist Mechanical Aptitude Tests." [3] The Assembly tests are bulky, cumbersome, and expensive. They are more or less inconvenient to administer and score. Stenquist has tried in these Mechanical Aptitude Tests to provide convenient simple tests which will measure the same abilities that are measured by the Assembling Tests. The tests consist of pictures of mechanical implements, tools, etc., and questions are asked which are likely to show evidence of mechanical insight. Two criteria [4] were held in mind in selecting the items and asking the questions.

1. Devices must be general in interest, and not pertain to very highly specialized trades. Common household articles that are of a mechanical nature are most apt to fall within the experience of every one.

[1] Stenquist, J. L. — *Measurement of Mechanical Ability*, Part II.

[2] Toops, H. A. — *Tests for Vocational Guidance of Children Thirteen to Sixteen*, p. 27.

[3] Published by the World Book Company.

[4] Stenquist, p. 64.

2. The question involved must be as mechanical as possible in its nature, involving a knowledge of, familiarity with, or understanding of the purpose, use, operation, construction, or reason for special size, shape, weight, material, etc., of the device in question.

Scoring is quite objective.

"A total of 173 questions, some expressed in terms of pictures to be compared one with another, and some in terms of printed queries referring to lettered pictures of machines and common mechanical articles were originally compiled into two tests I and II." In test I was placed only non-verbal material. The task is to determine which of five pictures "belong with, is a part of, or is used with" each of five other pictures. The total test has nineteen distinct group elements. The test is scored by counting the total number of items right.

Test II is divided into ten parts, each consisting of from five to seventeen questions. The first of these consists of nineteen pictures of mechanical toys, and each of these pictures has been cut into two parts. The task is then to find the missing part for each picture. Parts 2, 3, 4, and 5 consist of a series of questions relating to the mechanical properties of each of four lettered pictures of typical machines: an ordinary electric bell, a blower, a countershaft, a power drill press. The questions asked are, however, answerable by competent persons, even though they have not had direct experience with these particular machines, as they involve chiefly reasoning about and perception of mechanical relations. The reliability of this test is satisfactory (.70). Toops[1] finds that the Mechanical Aptitude (picture) tests correlate with general intelligence about .60 and it is estimated that they correlate with shop ranks about .60. As compared with the Assembling Tests the picture tests correlate much more highly with abstract intelligence and less highly with shop ranks. We may conclude therefore that they are inferior to the Assembling tests for the purpose for

[1] *Op. cit.*, p. 27.

which they were constructed, namely, as mechanical aptitude tests.

MACQUARRIE MECHANICAL ABILITY TEST

A newcomer in this field is the "MacQuarrie Test for Mechanical Ability." This is a paper-and-pencil test. For this test its author claims a reliability of .90, correlation with general intelligence of .22, and correlation with teachers' estimates of mechanical ability ranging from .48 to .80. On the basis of these correlations the MacQuarrie test would seem to be quite comparable to the "Stenquist Assembly Test" and distinctly better than the "Stenquist Paper Test" as a measure of mechanical aptitude. The tests in the MacQuarrie battery are:

1. Tracing
2. Tapping
3. Dotting
4. Copying designs, using a series of dots as guides
5. Location — locating dots in an abac of 81 letters
6. Blocks — a test of reading from a picture how many blocks a given block touches
7. Pursuit — following a line through a tangle of lines and indicating where the line ends

This test places a premium upon reaction time, speed, and manual skill.

REFERENCES

FILER, H. A., and O'ROURKE, L. J. — "Progress in Civil Service Tests," *Journal of Personnel Research*, 1 : 484–520, March, 1923.

Partially Standardized Tests for Junior Clerks. *Public Personnel Studies*, Vol. 3, no. 12, Ser. no. 27; pp. 346–371, 1925.

RUGGLES, A. W. — *A Diagnostic Test of Aptitude for Clerical Office Work.* TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 148, Teachers College, Columbia University, New York, 1924.

STENQUIST, J. L. — *Measurement of Mechanical Ability.* TEACHERS

College Contributions to Education, no. 130, Teachers College, Columbia University, New York City, 1923.

Thurstone, L. L. — "Standard Tests for Office Clerks," *Journal of Applied Psychology*, 3 : 248–251, Sept., 1919.

— "A Comparative Study of Clerical Tests." *Public Personnel Studies*, Vol. 1, nos. 2–6, 1923.

Toops, H. A. — *Tests for Vocational Guidance of Children Thirteen to Sixteen.* Teachers College Contributions to Education, no. 136, Teachers College, Columbia University, New York; 169 pp., 1923.

— "The Technique of Vocational Guidance as Investigated by the Institute of Educational Research," *The National Vocational Guidance Association Bulletin*, no. 6, pp. 86–92, Jan., 1923.

CHAPTER XIX

PREDICTION OF SUCCESS IN HIGH SCHOOL

Practically all cities to-day admit pupils to high school on the basis of having satisfactorily finished the work of the previous school, whether it be the sixth grade of the elementary school, the eighth grade of the elementary school, or the ninth grade in the junior high school. The elementary school decides whom it considers ripe to advance to the next unit in the system. Sometimes this is done on the basis of regular marks and sometimes by means of a special examination. But the high school does not concern itself particularly with the problem of admission. Children coming from other school systems usually are admitted to high school on the basis of credentials which they bring showing their previous school progress. This apparent carelessness in scrutinizing candidates for admission to high school is due to the present-day tendency to democratize secondary education. "The doors of the high school should be open to all" is a present-day slogan, and in the face of such public sentiment high school authorities do not care to challenge those who wish to enter high school. This attitude has a foundation in benevolent intentions, but like so many democratic notions, it operates somewhat blindly. The facts are that a large percentage (20%) of those who enter high school drops out before the second year and a large number of those who drop do so because they do not have the capacity to do the work. These facts indicate that the present policy is too liberal and even though we may admit that the high school should be open to all who wish to enter, schools could do more than they are now doing to insure that the pupils who come to them shall

be directed toward activities in which they will be successful. Placed in this light the problem becomes one of guidance rather than admission.

To handle this problem more rationally one needs to be able to predict success in high school curricula and courses. That is the problem which this chapter discusses. The general method of solution is indicated in the chapter on prognosis tests. In brief it is to find those factors measurable at the opening of the first year of high school which correlate most highly with success in the work of the high school, at the end of the first year or in later years. The difference between the point of view in this chapter and that of the chapter on prognosis tests is that now we are concerned not merely with tests but also with any records or data which may be available and significant.

School success is something that needs careful definition. Theoretically, perhaps, it should mean all that is of value which is gained from the school experience. Obviously this means success in school studies. One should include as well the gain made in character, socialization, physical fitness, and the like. These latter are intangible and can only be gauged subjectively. School success is usually measured by means of ordinary school marks, or perhaps more recently by standardized tests. A sufficient reason for this is that these indicia are the only ones which can be objectively judged and recorded. It is also true that these marks are measures of success in studies — which are, after all, the real business of the school — and hence ought to be a measure of school success. Furthermore, there is very strong presumption that scholastic success is correlated with these other less tangible yet valuable outcomes of school instruction. Using school success as synonymous with scholastic success, there are two measures to be used. One is the school marks at the end of a school semester or year or longer. The other is the length of time the pupil remains in school. These two are more or less correlated, however. It is

customary to use school marks as the measure of school success partly because they represent differences in the quality of school work, but also because they represent a measure that can be obtained rather soon after the original records are gathered, a desideratum of some influence to any one who is waiting for the results.

The problem then boils down to "How well can school marks in high school at the end of the first year be predicted from data available at the beginning of the year?" The following discussion follows the historical development of the answer to this question.

Miles [1] seems to have been the first to study the relationship between elementary school marks and high school marks, using the correlation coefficient as a measure of relationship.

In computing the individual's average elementary school grade all the marks received in his elementary school course were taken into account; similarly, all his high school marks were averaged to obtain his high school standing. The correlation was + .71.

Concerning this Kelley [2] says:

The fact that Miles deals with the average of all high school grades results in higher or lower correlations than would be obtained in dealing with first-year high school marks only, dependent upon which of the two following factors is the stronger:

(1) In general, as the time between testing is increased the correlation decreases; and as the second, third, and fourth years of the high school are more and more remote from the elementary school in time, it might be expected that correlation between the elementary school record and the first-year record would be greater than that between the elementary school record and the average of the entire high school record.

(2) A factor tending to offset this is the fact that the reliability of an average increases as the number of grades entering into it increases, and, to the extent that a grade represents native ability, the greater the number of grades averaged the greater the reliability of the measure. It is impossible

[1] MILES, W. R. — *Comparison of Elementary and High School Grades.*

[2] Cf. *Educational Guidance*, p. 77.

to say, *a priori*, which of these factors is the more important, but it is the author's opinion that the importance of the second factor has been, quite generally, undervalued, and might easily be the more important of the two.

Kelley [1] studied the value of elementary school marks, teachers' judgments, and specially constructed tests in predicting first-year average grade in high school.

The correlation of first-year average grade in high school with the various factors is as follows:

TABLE 98

CORRELATION OF FIRST-YEAR AVERAGE GRADE IN HIGH SCHOOL WITH VARIOUS FACTORS (FROM KELLEY)

7th grade average grade	.719
6th grade average grade	.728
5th grade average grade	.531
4th grade average grade	.624
Composite of elementary school grades	.789
Teachers' estimate of intellectual ability	.72
Teachers' estimate of conscientiousness	.62
Teachers' estimate of emotional interest in school work	.58
Teachers' estimate of oral expression	.63
Composite of teachers' estimate	.76
Composite of mathematics, English, and history tests	.51
Composite of all data	.89
Age	– .31

It will be noted here that the best prediction is obtained from elementary school marks, that the next best prediction comes from teachers' ratings, and the least valuable from the tests.

Fretwell [2] made a study of prognosis of success in grade 7, the first year of the junior high school. He found the following correlations with marks in the first year of the junior high school:

[1] *Op. cit.*, pp. 7, 11, 16, 71, 73.

[2] FRETWELL, E. K. — *A Study in Educational Prognosis*, pp. 39, 41, 1919.

TABLE 99

CORRELATION OF MARKS IN FIRST YEAR OF JUNIOR HIGH SCHOOL WITH VARIOUS TESTS (FROM FRETWELL)

	1916	1917
Thorndike Visual Vocabulary Scale A	.34	.33
Thorndike Reading Scale Alpha 2	.43	.40
Composition	.32	.35
Spelling	.32	.44
Trabue Completion B	.16	.22
Trabue Completion C	.31	.20
Woody Multiplication	.28	.33
Woody Division	.27	.30
Woodworth-Wells Opposites	.26	.36
Woodworth-Wells Easy Directions	.29	.23
Woodworth-Wells Mixed Relations	.17	.17
Average of eleven correlations	.29	.30
Composite of eleven tests	.57	.56
Marks in first six grades	.49	.49
Age	− .34	− .34

Fretwell finds that his best prediction is obtained from a composite of eleven tests, and that the marks in the first six grades are inferior to the tests as a basis for prediction. This is a direct contradiction to Kelley's findings. Kelley found that school marks give the best prediction; Fretwell found that tests give the best prediction. Note that Kelley's tests give only a slightly poorer prediction than Fretwell's tests (.51 against .56) whereas Fretwell finds that previous school marks are much less prognostic than Kelley finds them to be (.49 against .79).

Ross [1] has made an extensive study of the value of elementary school records as a basis for predicting high school success. His results are best reported by giving the correlations between average grade in the first year of high school and the variables

[1] Ross, C. C. — *The Relation between Grade School Record and High School Achievement.*

which he used in his study. The entire elementary school record was used in computing these correlations.

TABLE 100

AVERAGE CORRELATIONS BETWEEN MARKS IN THE ELEMENTARY SCHOOL AND MARKS IN THE FIRST YEAR AT HIGH SCHOOL (FROM ROSS)

	1916	1917	1918	1919
Reading		.35		
Spelling		.36		
Arithmetic		.52		
Geography		.50		
English		.59		
Fine Arts		.35		
History		.40		
Special subjects		.26		
Deportment		.35		
Effort		.37		
Days present		.00		
Simple average, all subjects		.60		
Weighted average of selected subjects	.68	.67	.56	.65
Age		− .36		

The correlations between composite scores obtained from grade school factors and average standing of pupils who remain in high school two or more years are .60, .64, .57, and .66, respectively, for four successive years. The factors in the elementary school record which Ross [1] found most significant for predicting high school success are:

Age at completing grade 8
Grade progress
English, grades 4–6
English, grades 7–8
Arithmetic, grades 7–8
Special subjects, grades 7–8
History, grades 5–6
History, grades 7–8
Effort, grades 7–8
Days present, grades 2–3
Days present, grades 4–6

[1] *Op. cit.*, p. 64.

Ross[1] also tried out three standardized tests which correlated with success in first year of high school as follows:

Grade school composite	.63
Terman Group Test of Mental Ability	.37
Thorndike-McCall Reading Test	.33
Woody-McCall Arithmetic Test	.40

Ross therefore supports Kelley in finding that elementary school records are more valuable than are tests in predicting high school success.

Ross[2] also studied the prediction of persistence in high school. The most significant factor was age, which correlated −.57 with persistence in school. By using age, English marks in grades 7 and 8, effort in grades 7 and 8, and days present in grades 4 to 6, Ross found multiple correlations with school persistence of .63, .63, .52, and .54 for successive years.

Flemming[3] in an extensive study of the correlation of various factors in the Horace Mann School gives data which throw some light on the problem. She gives the following coefficients of correlation with school marks. All of the data are obtained during the year in which the marks were given, so that hers is not a true prognosis problem. This undoubtedly works to make the correlation of the ratings higher than if they had been obtained the year before.

Combining tests and ratings Flemming obtained in the junior high school a multiple correlation of .8454 with school marks using teachers' estimates of intelligence, school attitude, energy, and chronological age. In the senior high school a multiple correlation of .7752 was obtained using the Terman Test, the Otis Test, rating for industry and energy and chronological age. Other combinations give nearly as high correlations with school marks.

[1] *Op. cit.*, p. 36. [2] *Op. cit.*, p. 66.

[3] Flemming, C. W. — *A Detailed Analysis of Achievement in the High School.*

TABLE 101

CORRELATION OF MARKS IN THE JUNIOR AND SENIOR HIGH SCHOOL WITH VARIOUS FACTORS (FROM FLEMMING)

	Junior High School	Senior High School
Terman Mental Group Test *B*	.60	.63
Miller Group Test *A*	.47	.49
Otis Higher Self-Administering Test *B*	.52	.32
Haggerty Reading Sigma 3	.53	.30
Chronological Age	− .29	− .30
Estimates of		
Health	− .21	.35
Energy	.45	.34
Intelligence	.80	.70
Industry	.69	.50
School Attitude	.74	.50
Emotional Balance	.44	.31
Leadership	.51	.30
Will and Persistence	.72	.41
Conscientiousness	.61	.31
Desire to Excel	.57	.46
Mean Teacher Rating	.70	.56

In this same connection Proctor[1] reports correlations with high school marks as follows:

TABLE 102

CORRELATION BETWEEN FIRST-YEAR HIGH SCHOOL MARKS AND INTELLIGENCE TESTS (FROM PROCTOR)

Binet I Q 1916–1917	.545
1918–1919	.487
Army Group Examination a	.343
Army Group Examination Alpha	.413

[1] PROCTOR, W. M. — *Psychological Tests in Guidance of High School Pupils*, pp. 11, 15, 16.

Jordan reports the following correlations with high school marks obtained by himself:

TABLE 103

CORRELATION OF HIGH SCHOOL MARKS AND GROUP INTELLIGENCE TESTS (FROM JORDAN)

Otis	.450
Army Alpha	.476
Miller	.476
Terman	.492

Jordan [1] also reviews the literature and gives the following as average or typical correlations with high school marks:

TABLE 104

AVERAGE CORRELATIONS OF HIGH SCHOOL MARKS AND GROUP INTELLIGENCE TESTS

	AVERAGE CORRELATIONS	NUMBER COEFFICIENTS	MAXIMUM REPORTED
Army Alpha	.38	26	.51
Otis	.49	14	.82
Terman	.47	9	.67
Miller	.56	1	

In order to facilitate comparison a summary showing the average or typical correlation with school marks found by different investigators is presented next.

From this table there are several discrepancies to be explained and some generalizations that can be drawn.

1. The correlation of elementary school marks and high school marks seems to vary considerably. Fretwell found as low as .49,

[1] JORDAN, A. M. — "The Validation of Intelligence Tests," *Journal of Educational Psychology*, 14: 348–366, Sept.; 414–428, Oct., 1923.

TABLE 105

RÉSUMÉ OF CORRELATIONS BETWEEN HIGH SCHOOL MARKS AND OTHER FACTORS

	Elem. School Marks	Tests in Elem. School Separately	Composite of Tests	Intelligence Tests	Ratings	Prognosis Tests	Age
Miles	.71						
Kelley	.789				.76	.51	− .31
Fretwell . . .	.49	.30	.56				− .34
Ross	.64	.37		.37			− .36
Flemming . . .				.51	.70		− .30
					.56		
Proctor				.52 Binet			
				.38 Group			
Jordan				.48			
Jordan's summary				.48			

Kelley as high as .789. When we know the conditions under which these two coefficients were obtained, a possible explanation presents itself for the discrepancy. Kelley obtained his data from the Horace Mann School. Fretwell did his work at the Speyer School. The Horace Mann School is a private school. All of Kelley's data were obtained in the school itself. Marks are probably assigned more accurately in the Horace Mann School than is customary in school situations. But of still more significance, since both the elementary marks were obtained in the same school and from only two or three teachers in each grade, they represent the same standard of marking throughout. The Speyer School was a public school in the City of New York. It was at the time of Fretwell's experiment an experimental junior high school for superior children. It drew from many elementary schools in the city. Many schools and many teachers were concerned in the elementary marks, and this fact would mean that the elementary

marks represent many standards of marking. Different standards always tend to attenuate coefficients of correlation toward zero, and there is no doubt that this force was operative in this case. The average correlation obtained by Ross may be taken to represent an average city high school situation where the pupils come from a few elementary schools which hold nearly the same standards in the system.

2. Tests of elementary school subjects correlate less with school marks than do elementary school marks. There seems to be something in elementary school marks over and above subject-matter achievement which correlates with high school success. What this is is not known distinctly, but it is probably something that corresponds to effort or industry or favorable attitude toward school work. Such character factors tend to become incorporated in school marks apart from pure achievement and these character factors tend to be more or less permanent, less permanent than ability, but stable enough to have some predictive value. The home background with its accompanying motive remains the same and early dispositions toward school work tend to persist. For these reasons elementary school marks may be explained to be superior to tests of elementary school subjects for prediction of high school success.

3. Intelligence tests are better than separate tests of elementary school subjects as predictors of high school success, but are not so good as elementary school marks. Intelligence tests probably predict better than separate subject-matter tests because they are more comprehensive. Fretwell found that a battery of subject matter and psychological tests gave a better prediction (.56) than is usually found with intelligence tests alone (.48).

4. Ratings of character and ability qualities have been found to give high correlations by Kelley [1] and Flemming.[2] Kelley found that teachers' estimates of intellectual ability correlated with

[1] *Op. cit.* [2] *Op. cit.*

school marks .72, and Flemming found .80 and .70 respectively in the junior and senior high school. Character qualities correlated slightly less. It is unfortunate that both of these studies have been conducted in the Horace Mann School. It may be that conditions there with superior teachers and an intimate acquaintance of teachers and pupils may give better results than a typical high school situation where the pupil load is much heavier. The unreliability of ratings in general leads one on general principles to discredit the predictive value of teachers' ratings. Even though these high correlations should be found in a public school situation one would hesitate to recommend the use of teacher ratings for predicting school success in general, because there is no guarantee that they have stability. A teacher in entering her ratings might be extremely careless. The same distrust is not felt concerning test results and school marks, the former because of the objectivity of the results, the latter because they are subjected to the censorship of pupils, parents, and school officials.

5. Probably not much more may be expected from tests especially constructed to measure school success than may be expected from our general intelligence tests.

6. Age correlates rather consistently with high school success around —.30.

7. By taking composites of various factors it is possible to get what may be called high correlations. Kelley [1] gets a multiple correlation of .89. Flemming [2] finds a multiple correlation of .85. Clem [3] in predicting success in Latin gets a multiple correlation of .84. These coefficients represent practically the maximum possible. The maximum correlation that one may expect (except by chance) between two variables 1 and 2 is represented by $\sqrt{r_{1I}\, r_{2II}}$. School marks are notoriously unreliable. If we estimate that the reliability of school marks is .70 (a generous

[1] *Op. cit.* [2] *Op. cit.*

[3] Clem, O. M. — *Detailed Factors in Latin Prognosis*, p. 49.

estimate) and of the composite of indicia used to predict school marks .95, we have as a maximum correlation $\sqrt{.70 \times .95} = .86$. From this it may be seen that by using four or five factors in a composite with a proper weighting one may predict school marks as well as the reliability of the school marks will permit. Before we can demonstrate that we can make a better prediction the reliability of school marks themselves will have to be raised.

8. Ross[1] is the only one to attempt to predict persistence through high school. Using age alone, he is able to predict length of time in school as accurately as it is represented by a correlation of .57. Probably economic and social conditions in the home are also important factors to use in predicting persistence through high school, as indicated by the studies of Van Denburg,[2] Holley,[3] Counts,[4] and others.

Attempts have also been made to predict success in particular high school studies. These results will not be reviewed in detail as they more or less duplicate or overlap the attempts to predict average school standing. Kelley predicted success in mathematics, English, and history, using the same data that he used to predict average success. Rogers[5] predicted mathematical ability, using tests already described in the chapter on Prognosis Tests.

Clem[6] made an extensive study of Latin prognosis, using 58 different factors. The criterion in this case was a specially devised Latin test and not school marks. The correlations with this criterion were as follows in one school:

[1] *Op. cit.*

[2] Van Denburg, J. K. — *Causes of the Elimination of Students in Public Secondary Schools of New York City.*

[3] Holley, C. E. — *The Relationship between Persistence in School and Home Conditions.*

[4] Counts, G. S. — *The Selective Character of American Secondary Education.*

[5] Rogers, A. — *Experimental Tests of Mathematical Ability and Their Prognostic Value.*

[6] *Op. cit.*, pp. 19, 20.

TABLE 106

CORRELATION OF A CRITERION OF LATIN ABILITY WITH VARIOUS FACTORS (FROM CLEM)

Briggs Analogies Alpha	.43	Civics	.40
Briggs Analogies Beta	.50	Drawing	.14
Thorndike Word Knowledge A	.30	Music	.11
Thorndike Word Knowledge B	.18	Physical Training	.20
Interpolation 1	.08	Teachers' Rankings on	
Interpolation 2	.34	Perseverance	.51
I Q	.48	Industry	.56
Age	−.38	Earnestness	.52
High School Attendance	.27	Nerve Stability	.32
Elementary School Attendance	.08	Orderliness	.57
Elementary School Marks		Self-Control	.56
Combined Average of Elementary Reading, Grammar, Composition, and Spelling	.33	Accuracy	.63
Arithmetic (Average of Two Marks)	.32	Right Attitude toward Criticism	.55
Combined Average of History and Civics and Geography	.28	Frequency of Help	.32
Combined Average of Penmanship, Music, and Drawing	.22	Promptness	.60
Combined Cooking and Sewing	.31	Ability to Do Independent Work	.62
Physical Training	.10	Social and Moral Attitudes	.47
General Estimate	.44	Minutes of Outside Help on Latin	−.04
Combined Average of 15 Marks	.41	Average No. of Movies a Week	−.17
High School Marks		Average No. Hours Sleep Daily	−.06
Semester Mark		Academic Interest as Shown by Plan after Graduation	.20
Latin	.84	Academic Standing as Shown by Plan for Life	−.13
English	.52	Time Spent Weekly Taking Music Lessons	.17
Biology	.70	Time Spent Weekly on Any Outside Language or School Subject	.12
		Time Spent Weekly in Work for Parents	−.01

TABLE 106 (*Continued*)

Average Time Spent in Study Daily		Pupils' Ranking of Latin on Basis of Importance . .	.16
English	−.25	Pupils' Ranking of Latin on Basis of Preference . .	.14
Latin	−.14		
Biology	−.24	Pupils' Ranking of Teacher on Basis of Preference .	−.04
Civics	−.25		
English + Biology + Civics . . .	−.30	Pupils' Participation in Extra-Curricula Activities .	.08

Note that outside of the teachers' ratings and the marks in other subjects the tests used by Allen give the highest correlations. Some one should determine the actual prognostic value of teachers' ratings as there is evidence that they have considerable correlation with later achievement. Note also the substantial *negative* correlations yielded by "average time spent in study daily."

Flemming [1] correlated the tests and teacher ratings not only against average school marks but also against marks in English literature, composition, grammar, mathematics, Latin, and history. The correlations are very similar on the whole to those with average school marks, although on the average lower. Concerning her findings she says:

The Terman and Miller tests show a correlation with the year's mark in English and mathematics about .10 lower than the coefficients with mean achievement. The correlation of Otis with English is about .06 smaller, while the association of scores on the Otis test with mathematics is almost precisely that with mean achievement. The correlation of Terman and history marks is .06 lower than with mean achievement. . . . Industry appears noticeably to be more of a factor in achieving success in Latin than in either mathematics or English, or achievement in general.

Jordan [2] finds a correlation with special subjects as follows:

[1] *Op. cit.*, pp. 124 ff.

[2] JORDAN, A. M. — "Correlations of Four Intelligence Tests with Grades," *Journal of Educational Psychology*, 13: 419–429, Oct., 1922.

TABLE 107

CORRELATION OF GROUP INTELLIGENCE TESTS WITH SEPARATE HIGH SCHOOL SUBJECTS

	AVERAGE MARK	MARK IN ENGLISH	MARK IN MATHEMATICS	MARK IN GENERAL SCIENCE	MARK IN HISTORY
Otis	.450	.412	.281	.407	.340
Army Alpha	.476	.391	.371	.480	.285
Miller	.476	.428	.342	.466	.147
Terman	.492	.424	.335	.508	.356

Ross [1] works out multiple correlation coefficients for predicting success in English, Latin, and mathematics from his elementary school data.

The following facts may be gleaned from a survey of the work done in predicting marks in special subjects.

1. The same factors which are used to predict general success in high school may be used to predict success in special subjects.

2. In general any one factor such as an intelligence test or mark in the elementary school correlates less well with a mark in a special subject than with the general average of subjects.

3. To predict success in a special subject one would use data that have something in common with the particular subject being predicted. To predict success in English in high school one would use marks in English in the elementary school rather than marks in arithmetic. Or one would use results from an arithmetic test rather than from a history test to predict success in algebra. However, this generalization cannot be indiscriminately applied. Ross,[2] for instance, finds that to predict Latin marks elementary school marks in English, arithmetic, and history are valuable. Allen [3] found that his battery of tests, designed to predict ability in Latin, also predicted ability in English and mathematics as

[1] *Op. cit.* [2] *Op. cit.*, pp. 28, 30 f.

[3] ALLEN, W. S. — *A Study in Latin Prognosis.*

well. High school subjects are specialized but the specialization is not clean cut — it is probably truer to emphasize the community between abilities in the high school subjects than their uniqueness.

4. Both Kelley and Ross found that they could predict ability in mathematics less well than ability in Latin. Evidently there is more specialization in mathematics than in language, although here again the difference is slight.

5. Clem's experience with teacher rating tends partly to substantiate the findings of Kelley and Flemming. Clem's correlations with ratings are slightly lower, but high enough to be significant in this connection.

The many factors which may be considered especially in their various combinations make the matter difficult to summarize. The following statement represents as nearly as possible the average of the findings.

School success may be predicted by a general intelligence test such as may be administered in an ordinary school period to the degree represented by a coefficient of correlation of .48. Separate tests of elementary school subjects correlate less well with high school success (.30–.40), but tests of elementary school subjects taken in combination are as satisfactory as a general intelligence test. Not much better results may be expected from a specially designed prognosis test for predicting average school success over a general intelligence test, but a specially designed prognosis test is slightly superior to a general intelligence test for predicting success in any one subject. Elementary school marks vary in their predictive ability. If all of the pupils come from the same teacher, the prediction is very high (.80); if from the same school, the prediction is somewhat lower (.70–.80); if from the same school system more or less homogeneously administered, the prediction of elementary school marks is about .65. Where the pupils come from a variety of elementary schools the predic-

tive value of elementary school marks is much lower. Taking the normal situation school marks are preferable to tests for prediction of success in high school. Teacher rating of personality traits, if carefully made, are also valuable. When well made, they can be depended upon to give correlations of over .50. In a special school situation such as the Horace Mann School these correlations from teacher rating run much higher. In any case, if carefully made, they undoubtedly provide a valuable addition to other data obtainable. Age correlates rather consistently about −.35. To predict success in any school subject, one follows about the same procedure that is followed in predicting general success. Tests and elementary school marks give slightly lower correlations. Tests and elementary school marks that have some community of material with the high school subjects give better prediction although the difference is slight. Ordinary intelligence tests and school marks predict success in the languages and history better than they predict success in mathematics, which seems to be somewhat (but not greatly) more specialized. Subjects like Latin, which are difficult, seem to be related to industry. There is evidence that subjects in which information is the main outcome, such as general science, are more closely related to general intelligence. The general rule is that prediction of general success is also about as good a prediction as can be made of success in any special academic subject.

And now just what should a school official use who wishes to predict school success? The answer seems to be that it depends on the type of school. A private school which retains its pupils from the elementary school through high school may predict success best by using the marks from the elementary school. Teacher ratings in such qualities as intelligence, school attitude, industry, and energy may also be used and a general intelligence test may provide a general supplement to the other indicia. A city high school in a city system which is well supervised so that

elementary schools all over the city hold the same standards may safely use elementary school marks to predict high school success but should supplement these by intelligence tests. Teacher ratings may also be used but less confidently. High schools, either public or private, drawing their pupils from a variety of elementary schools, should use a general intelligence test as the first basis for predicting success. Elementary school marks in such a situation are less valuable and probably add little to the information obtained from the intelligence tests. The same applies to teacher rating.

This discussion opens up more questions than it answers. Kelley's work gave the impetus to the movement, but his work is still unsurpassed. Many of the questions that his work raised are still unanswered. Except for the three or four studies mentioned in this chapter little has been done since Kelley's work in 1914. There is one important exception to this statement. The impetus given to the use of intelligence tests all over the country by the war has stimulated many to find the prognostic value of the general intelligence test. Much of this work has been of the copycat variety — one good study answers the question as well as several minor studies. At the present time apparently very little is being done in this field — and yet the opportunity is great. The work that has been done has uncovered many problems that are pressing for answer. These same problems have been more exhaustively studied with regard to prediction of college success. There is a very good reason for this. The persons best equipped to make investigations of this sort are college instructors and the college data are closest at hand. For a graduate student to take up this problem with regard to the secondary school is to break into new ground. High schools do not yet as a rule have research departments, and city research bureaus have to date devoted their attention to the problems of the elementary school.

What are some of the problems waiting to be answered?

1. What is the predictive value of a three-hour intelligence test as compared with the usual 30 or 40 minute test?

2. Can the predictive value of school marks be demonstrated to vary according to the homogeneity of elementary school standards as suggested in this chapter?

3. What is the predictive value of teacher rating of pupil characteristics in all sorts of school situations?

4. How well can success in the commercial courses of high school be predicted? This is an entirely unanswered question. All of the work to date on prognosis has been done in the academic high school. For purposes of guidance the matter should be investigated for the commercial courses.

5. How well can success in the domestic arts courses be predicted?

6. How well can success in trade courses be predicted?

7. The prediction of persistence in high school by correlation methods, using not only data from marks, tests, and ratings, but also economic and social status of the family.

REFERENCES

ALLEN, W. S. — *A Study in Latin Prognosis. Ref.* Ch. 17.

CLEM, O. M. — *Detailed Factors in Latin Prognosis. Ref.* Ch. 17.

COUNTS, G. S. — *The Selective Character of American Secondary Education*, SUPPLEMENTARY EDUCATIONAL MONOGRAPH, NO. 19. The University of Chicago; 156 pp., 1922.

FLEMMING, C. W. — *A Detailed Analysis of Achievement in the High School. The Comparative Significance of Certain Mental, Physical, and Character Traits for Success*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 196, Teachers College, Columbia University, New York, 1926.

FRETWELL, E. K. — *A Study in Educational Prognosis*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 99, Teachers College, New York; 55 pp., 1919.

HOLLEY, C. E. — "The Relationship between Persistence in School and Home Conditions," *Fifteenth Yearbook, Part II, National Society for the Study of Education.* Public School Publishing Company, Bloomington, Illinois; 119 pp., 1916.

JORDAN, A. M. — "Correlations of Four Intelligence Tests with Grades," *Journal of Educational Psychology,* 13 : 419–429, Oct., 1922.

— "The Validation of Intelligence Tests," *Journal of Educational Psychology,* 14 : 348–366, 414–428, Sept., Oct., 1923.

KELLEY, T. L. — *Educational Guidance. Ref.* Ch. 17.

MILES, W. R. — *Comparison of Elementary and High School Grades,* UNIVERSITY OF IOWA STUDIES IN EDUCATION, VOL. 1, NO. 1, 1910.

PROCTOR, W. M. — *Psychological Tests and Guidance of High School Pupils,* JOURNAL OF EDUCATIONAL RESEARCH MONOGRAPHS, NO. 1. Public School Publishing Company, Bloomington, Ill.; 125 pp., 1923.

ROGERS, A. L. — *Experimental Tests for Mathematical Ability. Ref.* Ch. 17.

ROSS, C. C. — *The Relation between Grade School Record and High School Achievement: A Study of the Diagnostic Value of Individual Record Cards,* TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 166. New York, 1925.

VAN DENBURG, J. K. — *Causes of the Elimination of Students in Public Secondary Schools of New York City,* TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 47. New York, 1911.

CHAPTER XX

PREDICTION OF SUCCESS IN COLLEGE

There is a more pressing need for predicting college success than for predicting success in high school. The same democratic philosophy that society applies to its public high schools is not held by the colleges, and quite rightly. Although one may grant the proposition that every individual should pursue his education as far as he wishes, the colleges do not pretend to provide educational facilities for all sorts of demands. American colleges avowedly devote their efforts toward providing "higher education," which means education on a level such that only those gifted with certain ability and preparation can profit by the offerings. American collegiate instruction in this sense is selective. With the growing demand for a college education there is a clamor for admission. Consequently selection to-day most prominently takes the form of a combing of the applicants and a *rejection of the unfit*. But selection has an opposite significance both in theory and in practice. Colleges positively aim to attract and select promising minds to profit by the advantages provided in the equipment, the faculty, and the communion with kindred minds, which they provide.

Those whose duty it is to sift the applicants for admission to college need a solution to the problem of predicting success so that those who cannot succeed may be denied admittance and those who can succeed may be taken in. The high schools are not uninterested in the matter of college entrance requirements, for these requirements influence high school education itself. In the past, the college preparatory curriculum of the high school

has been dominated by the college entrance requirements and this influence has extended to all the curricula of the high school. The colleges have practically controlled what has been taught in the high school both by requiring certain units in the high school course as a prerequisite for college entrance and by setting examinations in these subjects. This has made the high school curriculum static and has created a false goal in the students' minds as to the function of their high school education. Most schools would welcome a relief from this dominance by the colleges, and although relief might produce temporary unsettlement, it would eventually result in a rejuvenation of secondary education.

In 1918 Columbia College[1] first tried out an intelligence test on its entering class. A few years later data were published giving evidence that the intelligence test might be substituted for some of the other college entrance requirements. This was hailed as a blessing by all who were waiting for relief in this direction. Many have anxiously awaited the development of tests and corroborative data. This chapter attempts to evaluate intelligence tests as compared with the older indices as measures of college success. There are three main sources of data which have significance in the prediction of college success: (1) secondary school marks, (2) subject-matter examinations, (3) intelligence examinations. There are a few other less commonly used measures. The evidence concerning the value of each of these three groups of data will be presented in turn. More work has been done in attempting to predict college success than high school success, partly because colleges are more interested than high schools in selection of candidates for admission and partly because those trained in scientific investigation in education were college instructors who began their experimentation at home.

[1] *Columbia University Report on the Use of Intelligence Examinations in Columbia College*, p. 3 ff., 1922.

Secondary School Marks

MacPhail[1] has made a serviceable summary of the studies which have correlated high school marks and college marks. This table has been extended with additional data and is presented below:

TABLE 108

Correlation of College Marks with Secondary School Marks

Institution	Coefficient	Reported by
Arkansas, University of . .	.54	Terman
Carnegie Institute . . .	.29	Report Columbia Univ.
Columbia University . . .	.63	Terman
Columbia University . . .	.45	Report Columbia Univ.
Columbia University . . .	.35	Seashore
Cornell University . . .	.47	Report Columbia Univ.
Harvard University . . .	.56 .56 .58 .69	Beatley
Iowa, University of	.53	Smith, F. O.
Ohio, State University . .	.38	Report Columbia Univ.
Stanford University . . .	.54	Terman
Texas, University of . . .	.38 .65 .74 .69	Terman
43 Institutions (Engineering)	.29	Thurstone
Columbia University . .	.26	Wood
	.33	
	.15	
	.29	
Syracuse University . . .	.40	May, M.
University of Minnesota .	.63	Johnston, J. B.
Hawaii, University of		
Arts and Science . . .	.55 .54	Symonds
Applied Science . . .	.39 .29	
Stanford University . . .	.41 .52	Proctor-Cowdrey
Chicago, University of . .	.61	Scates
Brown University	.29	MacPhail

[1] MacPhail, A. H. — *Intelligence of College Students.* Warwick and York, Inc., Publishers, 1924.

An average of twenty-eight such coefficients is .47, which may be taken as a typical figure to represent the correlation of high school marks and college marks. This correlation should be compared with the correlation of .64 between elementary school marks and high school marks found by Ross.[1] This difference is partly due to the greater homogeneity of the college group, which always tends to reduce correlation. Secondly, the difference may be largely ascribed to the greater difference in standards of marking between different high schools than between elementary schools in the same school system. Thirdly, variations in the reliability of high school marks and college marks are always to be considered. Still another reason may be the greater specialization of college work which would naturally be less well predicted by high school work which, though general as we usually consider it, is particular in the sense that it differs from the specialized college courses.

These coefficients range from .15 to .74. Some of the lowest correlations are from engineering schools. A correlation of .29 is reported for the Carnegie Institute in Pittsburgh. Forty-three engineering schools using the Thurstone Psychological Examination have an average coefficient of .29. On the other hand, Harvard University, which has courses of an academic nature in the freshman year, reports correlations of .56, .56, .58, and .69. It would appear that high school marks predict college work of an academic character better than they predict specialized or technical courses. A direct comparison supporting this conclusion may be had from the writer's [2] experience in Hawaii. The following correlations were obtained between high school marks and college marks at the University of Hawaii:

[1] *Cf. The Relation between Grade School Record and High School Achievement*, p. 23.

[2] Symonds, P. M. — "The Significance of Intelligence Tests in the University of Hawaii," *School and Society*, 20: 601–606, Nov. 8, 1924.

TABLE 109

CORRELATIONS BETWEEN COLLEGE MARKS AND SECONDARY SCHOOL MARKS, UNIVERSITY OF HAWAII

	1ST SEMESTER	2ND SEMESTER
College of Arts and Sciences . . .	.55	.54
College of Applied Science	.39	.29

Here the effect of specialization on the size of the correlations is clearly evident. In general, high school marks may be expected to predict success in colleges of arts and sciences as well as represented by correlations of .50 and .60, whereas they predict success in technical schools and in colleges of applied science only as well as represented by coefficients of .30 to .40.

SUBJECT-MATTER EXAMINATIONS (COLLEGE ENTRANCE EXAMINATIONS)

MacPhail[1] has also reviewed the studies and tabulated the coefficients of correlation which show the relationship between subject-matter examinations and college marks. These are given with extensions in Table 110.

The average correlation between a composite of subject-matter entrance examinations and an average of college marks is .50. This would seem to be slightly higher than the average correlation between high school marks and college marks, which was .48. The evidence from institutions where the two sets of correlations have been computed on the same group of students is conflicting. Columbia College has data given by Wood[2] which permit a comparison of the marks on Regent's Examinations of New York State and high school marks. Regent's Examinations correlate with first-year college work .57 and .55, whereas high school marks

[1] *Op. cit.*

[2] Cf. *Measurement in Higher Education*, pp. 86, 87.

TABLE 110

CORRELATION OF COLLEGE MARKS AND SUBJECT-MATTER EXAMINATIONS

INSTITUTION	COEFFICIENT	REPORTED BY
Carnegie Institute (Ave. 6 tests SPEE)	.55	Report Columbia University
Columbia College	.43	Report Columbia University
Columbia College (C. E. B. Examinations)	.45	Seashore
Columbia College (Regent's Examination)	.57	Report Columbia University
Columbia College (Regent's Examination)		Seashore
Harvard University (Old Plan)	.41 .47	Beatley-Lincoln
Harvard University (New Plan)	.46 .50	Beatley
Ohio State (Ave. 6 tests SPEE)	.55	Report Columbia University
Smith College (New Plan)	.46	Report Columbia University
Smith College (Old Plan)	.40	Report Columbia University
Columbia College		
(Junior year college)	.47	Thorndike
(Senior year college)	.25	Thorndike
Columbia College (Regent's)		
(two-years scholarship)	.64	Wood
(Regent's)		
(first-year scholarship)	.57	
(Regent's)		
(second year first semester scholarship)	.40	
(Regent's)		
(first-year scholarship)	.55	

correlate with first-year scholarship .33 and .29. Just the opposite is found at Harvard, where the entrance examinations correlate with college success .41 and .47, while high school marks correlate with college work .56, .56, .58, and .69. It is difficult to reconcile these two contradictory findings.

The engineering schools, however, tell a rather consistent

story. Carnegie Institute reports[1] correlations with college success of .29, using high school marks, and .55, using a special examination. For Ohio State University the figures are .38 and .55, respectively. Apparently for specialized college marks special examinations are more valuable for prediction than the high school record.

Beatley[2] at Harvard finds that slightly higher correlations are found using the new plan (comprehensive) examinations (.46 and .50) than using the older plan separate examinations (.41 and .47). This is in accord with our knowledge that comprehensiveness in an examination always makes it more valuable for predicting success.

The Iowa Placement Examinations are in part objective examinations of subject matter and in part aptitude tests of an objective type. They will be described in detail later. They predicted college success at the University of Hawaii by an average correlation of .416, while at the Case School of Applied Science a composite of percentile scores of the Iowa Placement Examinations yielded a validity coefficient of .752. In general, higher correlations will result from using objective examinations than from using traditional type examinations.

Some of the same facts that are true for prediction of high school success are also true for predicting college success. School marks contain a record of a pupil's industry and effort during the preparatory course which is valuable as an indication of the student's probable attitude toward the work of the college. School marks represent different standards as they come from different schools, and insofar as a college draws from many different preparatory schools and preparatory schools of different types, these marks fall away in predictive value. Old-style subject-matter exami-

[1] "Report on the Use of Intelligence Examinations in Columbia College," 1922.

[2] Beatley, B. — "The Relative Standing of Students in Secondary Schools on Comprehensive Examinations and in College," *School Review*, 30: 141–147, 1922.

nations are notoriously unreliable and school marks have likewise been criticised on this score, but examinations and marks differ in this respect. Then again, different colleges have different standards of specialization — some specialize in courses with a professional outlook from the start; others are strictly academic. These factors in varying combinations can probably account for the discrepancies that exist between the predictive value of subject-matter examinations and school marks.

In general it may be concluded that for predicting success in specialized courses such as are given in a college of applied science, subject-matter examinations, carefully chosen, are better than high school marks. For predicting success in the arts and science courses in college high school marks are superior. High school marks correlate .50–.60 with college success in academic work. If we eliminate the two engineering schools from the correlations in the table in this section, entrance examinations correlate with college success something less than .50. The difference between high school marks and entrance examinations in predictive value for academic college work is slight but probably real. Corroborative data are supplied by Whitman [1] in connection with the prediction of separate courses in college. However, examination of an objective type may give better results than the traditional examination.

Psychological (Intelligence) Examinations

The amount of data on the predictive value of intelligence tests in college work is large. It would seem that almost every college in the country would soon have tried out at one time or another intelligence tests and reported the results. Toops [2] tells

[1] Whitman, A. D. — *The Value of the College Entrance Examination Board Examinations on Prediction of College Success*, 1926.

[2] Toops, H. A. — "The Status of University Intelligence Tests in 1923–1924," *Journal of Educational Psychology*, 17 : 23–36, 110–124, Jan., Feb., 1926.

us that in 1923–1924 60% of a random selection of colleges were using intelligence tests. MacPhail[1] made an inventory of all the coefficients reported at the time he wrote (July, 1923) and reported that the central tendency is between .40 and .45. Toops, in a more recent canvas of the situation, reports: "The median validity coefficient of 43 colleges which report their validity coefficients is .46." Taking these figures at their face value, it would appear that of the three — high school marks, subject-matter examinations, and intelligence tests — intelligence tests come off most poorly. But there is more to the story than this. If we compute the median coefficient of correlation by separate tests used, we get the following:

TABLE III

AVERAGE CORRELATION OF VARIOUS INTELLIGENCE TESTS WITH COLLEGE MARKS

Test	Jordan	MacPhail	Toops
Thorndike Intelligence Test for High School Graduates		.50	(.52 – .55)
Brown		.47	(.40 – .43)
Thurstone		.33	.40
Otis			.44
Terman		.48	(.44 – .47)
Army Alpha	.415	.39	.40

From this table it is seen that the predictive value of different examinations differs and that this predictive value is roughly correlated with the length of the examination. The Thorndike test, taking 2 hours 50 minutes of working time, has the highest validity, and the Army Alpha, requiring only 20 minutes of working time, has a low validity. This leads Toops[2] to conclude:

[1] *Op. cit.* [2] *Op. cit.*

> To one conversant with the magnitude of the standard error of estimate of small validity coefficients, it would appear that few of these colleges have tests long enough in time limits, or of the proper character, such as to yield minimally acceptable validity coefficients. Such validity will, however, in many cases compare favorably with alternative methods of old-type entrance examinations. A great need for improvement is evident. As indicated later, one big improvement will come by lengthening the time limits of the total examination, and in increasing the number of questions or test items accordingly. . . . We have shown in Table II [containing validity coefficients and summarized in table III of this book] that validity does increase with time. It is interesting to note that in the case of the seven colleges obtaining validity coefficients of .60 or above, none has less than an hour testing program, and only two have less than a two-hour program.

Toops also suggests that such tests as the Otis, Terman, and Army Alpha, some of which were originally devised for testing in high school, are too easy for use in predicting college success. A recent report by Thurstone [1] on the new American Council Intelligence Test for colleges shows that the test is yielding satisfactory prognosis of college work. He reports a correlation with college success of .54 in the University of Chicago and of .60 in the Case School of Applied Science.

It may be concluded that no single answer can be given as to the predictive power of intelligence tests. It all depends on the test. A short easy test like Army Alpha gives definitely poorer prediction than either school marks or subject-matter examinations. A longer and harder examination like the Thorndike Examination of the American Council tests compares well with school marks as a basis of prediction and is safely superior to subject-matter examinations.

Among the criteria for entrance, colleges are now regularly requiring "15 units." Formerly there was no such uniformity, but statistical studies of college entrance requirements have

[1] THURSTONE, L. L. — "Psychological Tests for College Freshmen," *Educational Record*, 6: 69–83, 282–294, Apr., Oct., 1925.

shown that 15 units is the mode, and as usual Americans hasten to adopt the mode. If this movement had been reasoned, one would expect such a requirement to be based on some finding that *amount* of high school preparation is highly correlated with college success and that 15 units is the minimum essential for doing good college work. As a matter of fact, there are only two studies relating to this question and these are quite negative in their conclusions. This is evidence of the irrationality of much of our administrative procedure. Matters of economy and convenience often overrule efficiency. May [1] found that units of high school work presented at entrance correlated .22 with honor marks achieved in the college course.

In his work at Minnesota Johnston [2] tried out quality of written English and proportion of advanced studies taken during the high school course as indicative of college success. Johnston says:

> The English themes have low predictive value with reference to general success in college work. The fact that the student selects advanced or elementary studies during his high school course has a certain predictive value, but this has not proved as important in the later groups studied as in the first one.

Hughes [3] has done the most work in interpreting ratings of character and personality traits as indicative of college success. His latest report indicates that trait rating may be a valuable supplement to the more objective measures as criteria of college success, although he presents no correlations and hence does not permit comparisons with the other indicia already mentioned. His charts show, however, that there is a "general tendency for high scholarship to go with high ratings and low scholarship to go with

[1] May, M. A. — "Predicting Academic Success," *Journal of Educational Psychology*, 14: 429–440, Oct., 1923.

[2] Johnston, J. B. — "Predicting Success or Failure in College at Time of Entrance," *School and Society*, 19: 772–776, June 8, 1924; 20: 29–32, July 5, 1924.

[3] Hughes, W. H. — "Personality Traits and College Success of High School Graduates," *California Quarterly of Secondary Education*, 1: 225–235, 1926.

low ratings. . . . The study indicates, without doubt, the importance of including trait ratings in the records which are to be used for predicting college success or failure." Hughes should have put the matter beyond the shadow of a doubt by an adequate statistical treatment of his results.

MacPhail [1] also experimented with character ratings at Brown University.

Each high school principal who has sent a freshman to Brown since 1916 has been asked to indicate his estimate of the student in the twelve qualities of mind given below by placing at the right of each a number designating the student's rank (5 for excellent, 4 for good, 3 for medium, 2 for poor, and 1 for deficient).

Intelligence	Initiative
Studious habits	Coöperation
Integrity	Leadership
Punctuality	Popularity
Neatness	Cheerfulness
Perseverance	Health

The correlations of these ratings with college success was .168.

Chambers [2] studied the predictive value of the Pressey X–O tests and found rather striking results. Pressey [3] some years earlier had intimated that character tests of this kind might prove useful in predicting college success, but did not provide convincing data. Chambers tried out the tests and found correlations of .54 and .46 with college marks. These correlations warrant considerable more work being done on these tests in trying to perfect them and make them generally available.

Four workers tell us that better results may be obtained by combining various indicia than by using them separately.

[1] *Op. cit.*

[2] Chambers, O. R. — "Character Trait Tests and the Prognosis of College Achievement," *Journal of Abnormal and Social Psychology*, 20: 303–311, Oct., 1925.

[3] Pressey, S. L. — "A Group Scale for Investigating the Emotions," *Journal of Abnormal and Social Psychology*, 16: 55–64, Apr., 1921.

May[1] found that although he can predict "honor points" in college by

intelligence with a correlation of	.60
high school average grade correlation of	.40
units of entrance correlation of	.22

by using all three the multiple correlation is .84.

Wood[2] used the Thorndike Intelligence Test, Regent's Examinations and secondary school marks. Singly they correlate with his college scholarship score as follows:

Thorndike Intelligence Score	.55
Regent's Examinations	.55
Secondary School Marks	.29

Combined they correlate with college success .6623.

Johnston[3] at the University of Minnesota used a specially devised intelligence test and ranking in scholarship in the high school senior class. The separate correlations with first semester college marks is

intelligence	.63
high school rank . . .	.50

Combined, these two factors correlate with college marks .67. This same correlation is .68 for the women alone and .72 for the men alone.

Symonds[4] found at the University of Hawaii that intelligence tests correlated with college success to the extent of .41, high school average to the extent of .52, combined .586.

In each of these four studies both the intelligence test and high school marks supplement each other so that taken in combination they predict college success better than either one singly.

[1] *Op. cit.*
[2] Cf. *Measurement in Higher Education,* p. 87.
[3] *Op. cit.*
[4] *Op. cit.*

TABLE 112

Comparison of Combined High School and Intelligence Ratings with Freshman Scholarship (taken from Johnston)

Combined Percentile Ratings High School Marks and Intelligence	Freshmen Marks									
	F−	F+	D−	D+	C−	C+	B−	B+	A	Total
96–100						1	2	2	1	6
91–95							4	1	1	6
86–90				1	5	2	4			12
81–85	1			1	3	7	8	1		21
76–80			1	3	11	3	3			21
71–75	1			6	10	4	4	1		26
66–70	2		4	7	11	4	4			32
61–65			5	10	9	4		1		29
56–60	2	2	4	10	13					31
51–55	3		7	9	10	3				32
46–50	4		6	10	9	1	1			31
41–45	6	1	12	7	6	2				34
36–40	2		8	8	6	2				26
31–35	7	2	15	5	2	2				33
26–30	15	3	9	3	2					32
21–25	6	2	4	5						17
16–20	10	3	4	5	2					24
11–15	7	1	5	1						14
6–10	3	2		1						6
1–5			2							2
	69	16	86	92	99	35	30	6	2	435

Correlation .67 ± .02

The scatter diagram taken from data of Johnston [1] obtained at the University of Minnesota illustrates the predictive value of a combination of high school marks (percentile values) and intelligence percentiles. Note, for instance, the prediction of the 46–50th percentile. Four obtained F−, 6 D−, 10 D+, 9 C−, 1 C+, and 1 B−. This is a wide variation — the prediction is not at all

[1] *Op. cit.*

close. Or note the range of those who receive F — three had a percentile rank of 6–10; one had a percentile rank of 81–85. Yet the prediction as a whole has some value. Of the 66 who had percentile ranks above 75, only 7 did less than average (C –) work. Of the 63 who had percentile ranks of 25 or below only 2 had a college standing as high as C – and none did better than that. Looked at from this point of view, a correlation of .67 has considerable significance.

In discussing this persistence of standing in high school to standing in college one can always get examples of youths who do not fit into this scheme of things. Tales will be told of boys who did extremely poor work in high school because of lack of interest, fast growth during adolescence, or simply stupidity, and yet were found some years later in the upper five of the class in law school. Newton and Edison are usually referred to before such a discussion proceeds very far. Such cases should be discredited unless the testifier has *first-hand* evidence. When such cases are found, they really prove the general rule of correlation. The very fact that they are noted and carried in mind as striking examples proves that they are exceptional cases. Also it is true that some teachers never know real ability when they see it. Marks and intelligence tests can be depended to give results similar in the long run to the scatter diagram.

What, then, shall be concluded concerning the prediction of college success? Perhaps the first conclusion is the statement last to be made: That no one factor predicts college success adequately and that any criterion of college success should be the composite of several objectively measurable factors. Neither high school marks nor intelligence tests alone can be depended upon to give an adequate prediction of college success even under the most favorable circumstances.

High school marks are probably the most valuable single measure of college success of the academic type. They may be

depended to give correlations of from .50 to .60 even in their present unsatisfactory state. Johnston suggests that instead of percentage or letter high school marks, percentile standing in the high school class be used. This would surely help in eliminating the incomparability of marks from different schools. Such percentile standings, however, would be difficult to obtain. School marks, as has been said before, contain the results of both ability and the use to which ability has been put, and both of these correlate well with future scholastic success.

In technical schools or colleges of applied science high school marks are not so predictive.

Running close up on high school marks are intelligence tests. The Thorndike Intelligence Examination for High School Graduates and the new American Council Tests can be depended upon to yield average correlations of over .50 and even as high as .60 in favorable circumstances in academic courses. There are great differences between tests in predictive power. Length of test seems to be a potent factor — the most valid tests are the longest. Level of difficulty and selection of material are also important.

Least satisfactory for prediction are subject-matter examinations, at least in academic courses. On the average they yield correlations of under .50. Comprehensive examinations yield better results than separate examinations. In the technical schools selected examinations are more useful for purposes of prediction.

Means of Improving Prediction

Probably the first way of improving prediction is by making the criterion — college marks — more reliable. One cannot predict more accurately than the criterion which one is trying to predict is reliable. Wood[1] finds that .69 is the correlation of first and second semester marks in Columbia College; Symonds found

[1] *Op. cit.*, p. 133.

.73 for similar measures in Hawaii.[1] .70 may be taken as a typical figure. The maximum possible correlation, as explained in the last chapter, is $\sqrt{r_{1I}\ r_{2II}}$. Accordingly, validity coefficients around $.83 = \sqrt{.70}$ are the maximum one can expect in the way of prediction at the present time with college marks as reliable as they are. Many writers have mentioned this. Wood has done excellent constructive work in Columbia College in getting various departments to adopt more objective forms of testing in order to raise the reliability of marking. Concerning this problem, Wood[2] says:

> We have shown thus far that the intelligence examination predicts college success very nearly as well as college success, as expressed in terms of the best available index, predicts itself from one semester to another or from one year to another. Since a test cannot predict a criterion better than the latter forecasts itself, all hope of improvement in the intelligence test's prediction depends upon improvement in the reliability and significance of the criterion. We have not found all the causes of the unreliability of our measures of college success, nor have we any very hopeful suggestion for the elimination of some of the causes that we have definitely located, such as the obvious inequalities (both qualitative and quantitative) between various college courses of the same number of credit points and hours of recitation. But a very large part of the unreliability, that due to the indefinite and variable bases of grading and to wrong principles and poor technique of constructing and administering examinations and tests in courses of study, is remediable.

A second means of improving prediction of college success is by making the high school marks more objective and more comparable. School marks are already our most significant item. Anything that can be done to make them more significant will be a great gain. Wood has for several years been campaigning for better secondary school marks. Before this can be done secondary schools must be shown in a definite way how to give

[1] *Cf.* "Significance of Intelligence Tests in University of Hawaii."
[2] *Op. cit.*, p. 139.

better marks. It is the hope that this book will help in giving this concrete aid. Part of it lies in deciding just what a school mark means and what it should include; part in using objective tests; and part in using standardized tests. With regard to this means of improving college prediction let Wood[1] speak for himself:

It has recently been widely noticed that very little use is being made, and that very little use can be made under the present system, of the previous school records that students present. In the great cosmopolitan colleges which draw their students from all parts of the country, College Entrance Board Examinations do predict college success better than the average high school grades. (This was true at Columbia but not at Harvard.) But wherever a large number of students goes from one first class high school to one college, the high school record has been found to be a very much better prediction of college success than the College Entrance Board Examinations. This would seem to indicate that while our lack of faith in high school records is justified, our lack of faith in the value of the high school experience is not justified. More positively, it indicates that if we had more accurate and more complete records of the high school experience, entrance examinations might not be necessary. Our recent experience with new methods for the measurement of specific achievement is sufficient to convince us that it is now possible to secure objective, reliable, significant, and comparable measures of specific achievement throughout the whole educational ladder and it is only natural that there should be a strong demand for the realization of these much hoped for possibilities.

A third way of improving prediction of college success is by refining and lengthening intelligence tests. Thorndike has profited by Wood's studies and has reconstructed the Intelligence Test for High School Graduates by omitting certain tests shown to have lower validity, lengthening tests having good validity, and changing the form of some tests to make them more objective. The American Council project for improving intelligence tests for

[1] Wood, B. D. — "The College Curriculum and Vocational Guidance," *School and Society*, 2: 508–514, April 25, 1925.

use in college is excellent. Thurstone,[1] in describing this project, says:

> It is planned, further, to select the best tests every year for use the following year and to eliminate those tests which are not universally satisfactory. These eliminations will make room for the new test material which is submitted on the basis of preliminary experiments by individual psychologists. In this manner the selective value of the test program will be improved every year.

School marks can be improved only by improving the technique of every person using them, but a standardized test can be improved by the painstaking analytical work of a person or a group of persons and its resulting form will be everywhere available. Consequently there is more hope for immediate progress in the revision and refinement of intelligence tests than in the attempt to improve school marks.

Prediction of Success in Separate Courses

The foregoing discussion concerns the prediction of success in the college course as a whole, considering the average of all work taken. There is also interest in the prediction of success in separate courses. In the strictly academic careers, specialization is not usually expected to a marked degree. But one of the outcomes of almost any college course should be a selection of a special field of interest. In the technical and professional courses a specialization of work and interest is expected from the start.

Three investigators have made intensive studies of prediction in separate courses, using different criteria.

Gowen and Gooch [2] have made elaborate correlation studies of the relation of high school marks to college marks in the University of Maine. The college marks cover a period of eight years

[1] *Op. cit.*

[2] Gowen, G. W., and Gooch, M. — "The Mental Attainments of College Students in Relation to Previous Training," *Journal of Educational Psychology*, 16: 547–568, Nov., 1925.

from 1909 to 1917. Only the records of 927 students who finally graduated from their college course were included, which selection tends to make the coefficients lower than they otherwise would be by taking unselected college freshmen. Approximately 300 high schools are represented in the data so that the phenomenon of difference in standards of marking between schools should be as severe as will ever be found.

TABLE 113

COEFFICIENTS OF CORRELATION BETWEEN HIGH SCHOOL MARKS AND COLLEGE MARKS FOR SEPARATE SUBJECTS

Data from University of Maine

High School Marks	College Marks						
	English	Chemistry	Algebra	Analytical Geometry	Elementary French	Advanced French	German
English	.284	.209	.220	.276	.403	.389	.190
Chemistry	.198	.193	.335	.311	.256	.181	.250
Algebra	.184	.211	.407	.384	.427	.259	.313
Geometry	.190	.227	.409	.337	.135	.031	.279
Latin	.229	.228	.259	.297	.137	.316	.387
El. French	.252	.283	.408	.375		.411	.373
Adv. French	.312	.257	.329	.398		.350	.434
Average	.236	.230	.338	.340	.272	.277	.318

A scrutiny of the table shows, first, that the coefficients of correlation are lower than those involving average marks. The two factors mentioned above account for this in part, but my own estimate is that these factors are not sufficient causes and that the correlations between separate courses really are lower. Second, it is interesting to note that the highest correlations are not always between those of the same name. The high school course that correlates most highly with college English is advanced French; the high school course that correlates most highly with college chem-

istry is elementary French. Analytical geometry is best predicted by advanced French, and college elementary French by high school algebra. The only thing that these anomalous results indicate is that the mechanical or statistical factors which make for differences between the correlations outweigh the intrinsic differences between the subjects themselves in predictive value. It would seem that the factors of ability which make for excellence in general outweigh the special abilities needed for success in particular subjects. These low correlations would then be interpreted as being caused by factors of unreliability and selection rather than an intrinsic specificness in the function. Third, the difference between the average coefficients is probably due to differences in the reliability of college marks rather than differences in the specificness of the subject. If specificness were the factor in determining the average size of the correlation, English would have the highest average because it is least specific and analytical geometry the lowest because it is the most specific. The exact opposite is the case. Accordingly, the differences must be interpreted as due to the reliability of the college marks and the variations in standards in high school courses between different high schools.

Whitman[1] studied the relative value of school marks and traditional subject matter examinations for predicting separate college subjects. He had access to records of the College Entrance Examination Board and worked his data up in great detail for separate subjects and separate schools. Unfortunately he erred in assuming that each subject predicts best that subject which has the same name. The work of Gowen and Gooch shows the fallacy in such an assumption. The following correlations are between high school subjects and the college marks in corresponding subjects. Each coefficient in the following table is the average of several others, and represents several hundred cases.

[1] *Op. cit.*, p. 44.

TABLE 114

CORRELATION BETWEEN COLLEGE MARKS AND SUBJECT-MATTER EXAMINATIONS AND SCHOOL MARKS FOR SEPARATE SUBJECTS (FROM WHITMAN)

ENTRANCE EXAMINATION	EXAMINATION AND COLLEGE MARKS	SCHOOL MARKS AND COLLEGE MARKS
English (Comprehensive)	.39	.28
English (Old Style)	.19	.43
English (All)	.24	.39
French (Comprehensive)	.47	.33
French (Old Style)	.43	.39
French (All)	.45	.36
Algebra	.28	.43
Geometry	.22	.54
Mathematics (Old Style)	.25	.49
Mathematics (Comprehensive)	.33	.37
Mathematics (All)	.27	.45
History	.26	.31
Science	.41	.43
Average	.29	.40
Average (Comprehensive)	.40	.33
Average (Old Style)	.29	.44

The following points may be noted from Whitman's data: (1) School marks are more predictive of success in particular subjects than are subject-matter examinations. This conclusion has already been reached in considering the prediction of general collegiate success. (2) The predictive value of both examinations and school marks is lower for separate subjects than for a general average of subjects. (3) The comprehensive examination gives higher correlations than the old-style examinations. (4) The school marks of those who tried comprehensive examinations are less predictive than for those who tried the old-style examination.

These college entrance examinations studied by Whitman were

of the subjective type. A study by Stoddard [1] describes the predictive value of subject-matter examinations of an objective type. These examinations are known as the "Iowa Placement Examinations." They have been constructed in two groups. One, called the *training* group, is essentially achievement tests; the other, the *aptitude* group, is prognostic tests and may be given to students who have never taken the subject in question. The following tests are available.

Chemistry Training	Chemistry Aptitude
English Training	English Aptitude
French Training	Foreign Language Aptitude
Mathematics Training	Mathematics Aptitude
Physics Training	Physics Aptitude
Spanish Training	

TABLE 115

CORRELATION OF COLLEGE MARKS AND IOWA EXAMINATIONS

1st Semester Grades in College	Tests Used				
	Iowa H. S. Content Exam.	Iowa Comprehension Test	Thorndike Intelligence Exam.	Placement Exam. T = Training A = Aptitude	
Chemistry . .	.238	.335	.270	T	.341
	.379	.328	.407	A	.401
English . .	.393	.474	.418	T	.615
	.453	.439	.359	A	.422
French . .	.416	.306	.251	T	.626
	.371	.343	.179	A	.428
Mathematics .	.496	.435	.229	T	.528
	.295	.255	.173	A	.407
Average .	.380	.364	.286		.471
			av.	T	.528
			av.	A	.415

[1] STODDARD, G. D. — *Iowa Placement Examinations*, UNIVERSITY OF IOWA STUDIES IN EDUCATION, VOL. III, NO. 2, pp. 75–78, 1925.

The results of this work at Iowa are remarkable. Objective examinations in special subjects are much superior to the subjective examinations of the College Entrance Board. The Iowa High School Content Examination is a comprehensive objective examination testing (1) English, literature, etc., (2) mathematics, (3) science, (4) history and social sciences.

This gives an average coefficient of .380, which is very nearly as good as school marks .40 reported by Whitman. The special placement examinations give a still higher figure, .471. The training series yield the highest average correlation with their subjects (.53) and the aptitude series yield the lowest correlations (.41).

Stoddard [1] reports correlations from the University of Arkansas which tend to substantiate these figures.

TABLE 116

CORRELATIONS OF COLLEGE MARKS AND IOWA PLACEMENT EXAMINATIONS IN UNIVERSITY OF ARKANSAS (FROM STODDARD)

To Predict College Grades in	Iowa Placement Examinations		
	Training Series	Aptitude Series	Combined
English 1st semester . .	.563	.643	.608
2nd semester . .	.467	.329	
3rd semester . .	.474	.424	
Mathematics 1st semester	.466	.696	.562
2nd semester	.339	.366	
3rd semester	.490	.422	
Average	.467	.480	
Average 1st semester	.514	.670	.585
Average 2nd semester	.403	.348	
Average 3rd semester	.482	.423	

[1] *Op. cit.*, pp. 88, 89.

In the University of Arkansas the aptitude series was slightly superior in prognostic value to the training series, and both series show considerably higher predictive value than the old-style College Entrance Board examinations.

Comparatively few studies report the predictive value of intelligence tests for separate college studies. Stone [1] reports an average correlation of .307, using Army Alpha to predict standing in separate studies at Dartmouth. Perrin [2] reports an average coefficient of .425 in predicting ten separate college studies, using a specially devised intelligence test. Stoddard, in the study above reported, gives correlations using the Thorndike Intelligence Examination against first semester marks as follows:

Chemistry	.270
	.407
English	.418
	.359
French	.251
	.179
Mathematics	.229
	.173
Average	.286

A general intelligence test does not appear to be as valuable in predicting results in separate studies as do subject-matter examinations of the objective type. Intelligence tests apparently predict subjects having a specialized subject-matter, such as mathematics or a foreign language, less well than subjects which

[1] STONE, C. L. — "The Significance of Army Alpha in College," *Journal of Educational Psychology*, 13: 218, May, 1922.

[2] PERRIN, F. A. C. — "Freshman Psychological Examination, in the College of Arts and Sciences at the University of Texas," University of Texas Bulletin No 2445, Dec. 1, 1924.

are more general or familiar. The matter of prognosis of ability in separate subjects is very complex and the factors concerned are subtle and little understood, so that perhaps premature generalizations should not be attempted.

The high correlations found from the Iowa Placement Examinations need some explanation. The following critical remarks of Thurstone[1] are most suggestive.

If tests are selected merely on the basis of their correlation with average scholarship records, in all freshman subjects, it would be found that the psychological tests would be loaded with certain linguistic forms such as Completion, Opposites, and Analogies, while the more exact quantitative tests would find no place in the program. . . . The explanation of this fact is that the very tests which differentiate ability for particular subjects, such as German and French, Mathematics and Physics, et cetera, are the very ones which give, in general, the poorest correlations with average scholarship. If our objective is to predict average scholarship, then, in general, the formal linguistic tests are the most desirable. But if our objective with the psychological tests is to ascertain aptitudes for particular subjects, then certainly the discriminating value of the ordinary linguistic forms of intelligence test shows a poor performance. Since it is desirable to separate the student's performance into separate categories by means of separate tests in such a way that the profile may throw some light on the student's aptitudes, it is desirable to obtain, in a test program, not only these linguistic forms but also those tests which are found to have the highest discriminating value for different kinds of talents. For example, tests in arithmetic, estimating, and reasoning give, by far, the best discriminating value for performance in physics and in mathematics, but they give the poorest correlations with average freshman scholarship. On the other hand, tests such as Completion and Artificial Language give acceptable correlations with average freshman scholarship, but they have practically no differentiating value for aptitude in exact sciences. The same kind of comparison can be made for other tests with reference to other special aptitudes.

There is another problem concerned with the distinction between content examinations and psychological tests that deserves careful consideration. If our sole objective is to predict scholarship grade in a particular subject or average scholarship in a particular course of instruction, then clearly the

[1] THURSTONE, L. L. — "Psychological Tests for College Freshmen."

best diagnostic value will be obtained by means of tests which are very similar to the content of the course. If, on the other hand, we want a test to measure aptitude, as far as possible free from the immediate efforts of high school courses in the particular subject, then we have recourse to psychological tests proper which are designed so as to be relatively unaffected by particular courses of instruction. As a consequence, the correlation between scholarship grades in a particular subject and the score in a psychological test will necessarily be lower than the similar correlation between scholarship and the content examination. Here, again, the choice between the two types of examinations must depend on the objectives which should be clearly stated before making a choice of tests.

The Iowa Placement Examinations are perhaps the best form of objective content examination available. They certainly should be highly recommended and extensively tried for measuring a student's high school training in each particular subject. To give the whole program of the Iowa Placement Tests would, of course, require considerable time, and the best arrangement undoubtedly is to induce each department to devote its first class hour of the introductory course to an objective content examination such as the Iowa Placement forms. These examinations should give a good measure of the preparation of the student in the particular subject.

The Iowa Aptitude Tests constitute a cross between the objective content examinations which measure training and the psychological examinations which measure aptitude as far as possible independent of training. The Iowa Aptitude Tests do not depend so much on training as regular content examinations and college entrance examinations, but most of them are more influenced by training than the typical psychological examinations. Now, if we study the predictive value of the tests, we should expect to find the highest correlation between scholarship and content examinations, while the lowest correlations would probably be found for the psychological tests in which subject matter is intentionally eliminated. Very likely the best results in the long run will be obtained by a combination of tests measuring training and aptitude. This may conceivably be accomplished entirely by a combination of psychological tests and content tests or by a combination of both types in the same general form, as illustrated by the Iowa Aptitude Tests.

Determining Fitness for College

It goes without argument that not every one should be permitted to enter college. There must be some restrictions placed

on entrance, some standard of qualifications which will insure that the student body is ready and able to profit by college instruction.

McKown [1] states that

> in 1913 only 5 out of 314 colleges kept the enrolment of students to certain limits. In 1922, 19 of the colleges stated or implied that enrolment was limited. . . . Thus it is seen that limitation is not a fancy, but an actuality making its appearance in definite practice. The colleges are beginning to limit in two ways: First, by requiring higher intellectual attainments, and second, by devising definite procedure and providing the mechanism which will evaluate the other potentialities of the applicant. The movement is only in its infancy. Another decade or so will undoubtedly show practically all colleges using some specific means of limiting enrolment.

Whether or not limitation of enrolment becomes widespread, it is certain that there is a growing tendency to be more stringent in admitting to college.

A typical requirement for admission to college to-day is:

1. The applicant must be 16 years of age or over.
2. The applicant must submit a certificate of moral character.
3. The applicant must submit a certificate of health.
4. The applicant must have completed successfully (or with an average grade of more than a certain %) 15 units of high school work.
5. The applicant must undergo certain tests of his ability or preparation which may be either:
   a. Traditional subject-matter examination (old style)
   b. Traditional subject-matter examination (comprehensive)
   c. Objective subject-matter examination
   d. Intelligence test
   e. Average of marks obtained in high school

[1] McKown, H. C. — *The Trend of College Entrance Requirements*, pp. 112, 115, 1913.

The first three of these do not concern us particularly. They are safeguards which have little correlation with the work to be expected in college. The age requirement safeguards the social adjustments of the college youth more than his mental adjustments. Probably a more significant safeguard as far as the mental adjustments are concerned would be an *upper* age limit. The certificate of moral character and physical fitness is a requirement that is intended to protect the college community from disciplinary and health problems. With knowledge of the unreliability of testimonials or certificates of this kind one questions their usefulness — however, they are probably better than a complete ignoring of the problem and tend to prevent the acceptance of abnormal cases.

The requirement of 15 points of high school work is intended to obtain adequate scholastic preparation for college work. That this requirement is necessary or even helpful is dubious. There has been no extensive study of the matter of prerequisites for college work. Such critical experiments can only be made when some progressive college is willing to accept men who have inadequate high school experience but whose other qualifications are satisfactory. Correlations seem to indicate that prerequisites are not so important as we have been led to believe by the almost unanimous practice of the colleges. The study of Gowen and Gooch shows that college English can be better predicted by high school French than by high school English. Of course this does not say that college English does not require certain standards of preparation, but does imply that success in college English is much less dependent on high school preparation than hitherto supposed. We need extensive experiments to give evidence of what would happen in the case of a boy with good intelligence but incomplete high school training. Terman [1] cites certain isolated examples which show how the wind blows.

[1] TERMAN, L. M. — "Intelligence Tests in Colleges and Universities," *School and Society*, 13: 481–494, April 23, 1921.

J. P. M. left school at the end of the fifth elementary grade. Alpha score, 157, which is equal to that of the average Yale freshman. His record thus far is as follows (at Stanford):

	Units	*Grade*
Transportation	5	B
English Composition	1	C
Citizenship	4	B +
Spanish	5	B +
Elementary economics	5	B
English, extemporaneous speaking	2½	C +
U. S. History	5	B
English, public speaking	2½	B

E. H. A. Two years of high school work. Alpha score 170. Grades: 3 units, elementary accounting, A; 5 units, elementary Spanish, B –; 5 units, European history, B –.

It is quite probable that colleges could be much more lenient than they now are in requiring certain prerequisites in the way of a high school course, a certain number of units, and certain prescribed courses, without lowering the quality of their work, provided they use other safeguards to guarantee the requisite ability of the student body.

Of all the indices of ability to do college work, marks in the high school course are the most significant. They are also the easiest for a college to obtain. Colleges should use the quality of work done in high school as the first index of college ability. Johnston's suggestion that percentile ranking in the high school graduating class be used instead of letter marks has the merit of eliminating one of the attenuating factors in high school marks, but these percentile ranks would probably be difficult to obtain. Harvard's scheme of admitting any student standing in the upper seventh of his high school class is significant and the results are awaited with much interest. If found feasible, this method of determining the qualifications of applicants would also have the credit of being the easiest and simplest to administer.

School marks, however, should not be relied upon alone, but should be supplemented by intelligence tests. All the studies that have tried out both criteria find that they tell more together than either one does alone. A good intelligence test, such as the Thorndike or the American Council test, is much more valuable for predicting general academic college success than subject matter examinations, even when the latter are of the objective type. Of this Thorndike[1] says:

This 0.60 (correlation between psychological score and college mark) is a closer prophecy than any three-hour examination of the traditional type has ever been shown to give. It is, in fact, as good as the correlation obtained from using the total information from 16 to 20 hours of examination of the traditional type. Much more important than this, however, is the fact that by adding it to a system of content examinations, or a system of ratings of the candidate by the school where he comes, or a system of inspections of that school by the college, the selection is improved. A competent psychological examination, that is, gets at factors which have prophetic value and which no other existing means gets at in the same way. It may or may not be desirable in certain cases as an alternative to these various systems, but it certainly improves the selective action of any one of them.

Its value as an adjunct to a system of crediting school work, in fact, deserves special mention. Such systems, beneficent as they are, have the necessary weakness of overweighting educational regularity, docility, and good intent. The boy or girl who does the regular thing in the regular way and wins the moral approval of his teachers may slip along into the university in spite of a lack of essential brain power needed to profit by a college or professional course. You all know such cases. A competent intelligence examination discovers them almost without fail. On the other hand, the boy or girl whose schooling has been irregular, or whose abilities are specialized, or whose conformity to teachers' edicts is imperfect, finds himself delayed or even barred in his progress on to college. A competent intelligence examination differentiates these irregulars into the strong, who are just the element our colleges need, who have not met past requirements by reason of their weakness.

[1] THORNDIKE, E. L. — Report of U. S. Commissioner of Education, 1920, p. 14.

Technical schools and colleges giving professional courses can rely much less on high school marks, it seems, for prediction of success. They must rely on specially devised subject-matter tests and secondly on intelligence tests.

Needed Research for the Prediction of College Success

Before suggestions as to needed research are made I should like to emphasize the importance of utilizing the research already done. We know a great deal to-day about the prediction of college success which is not being utilized. What we know, for instance, if applied, would spell the immediate extinction of the College Entrance Board Examinations, at least in their present form. The use of intelligence tests would also be greatly extended if the colleges of the country were only aware of the facts of recent research.

Some suggestions for research are:

(1) The prediction of success in technical schools, schools of education, schools of business, schools of journalism, schools of medicine, etc. Examinations are of greater importance in these schools than in regular academic colleges and tests need to be devised and tried out in these professional schools.

(2) The value of character ratings as criteria of success.

(3) The value of interest questionnaires and other similar devices for prophecy of college success and for differential guidance.

(4) Means for standardizing and equating high school records.

(5) Means for making school marks, both high school and college, more objective and more reliable.

(6) Improvement of intelligence tests. Before any one attempts work on this problem he should familiarize himself thoroughly with what has already been done with intelligence tests and attempt to coöperate with present movements aiming

toward improvement. Information concerning current undertakings may be procured from the American Council on Education, 26 Jackson Place, Washington, D. C.

(7) Development of objective subject-matter examinations. The "Iowa Placement Tests" deserve a thorough trying out along this direction.

(8) Development of aptitude or prognosis tests.

(9) Relative predictive power for various separate courses of subject-matter tests, aptitude tests, and general intelligence tests.

(10) The effect of a limited high school course on college success.

(11) The question of the necessity or desirability of prerequisites for various college subjects.

REFERENCES

ALLEN, W. O. — "College Admissions," *School and Society*, 14 : 235–240, Oct. 1, 1921.

ANDERSON, J. E. — "Intelligence Tests of Yale Freshmen," *School and Society*, 2 : 417–420, April 3, 1920.

BEATLEY, B. — "The Relative Standing of Students in Secondary School on Comprehensive Examinations and in College," *School Review*, 30 : 141–147, Feb., 1922.

BREITWIESER, J. W. — *The Thorndike College Entrance Tests in the University of California.* University of California Press, Berkeley; 17 pp., Jan., 1922.

BRIDGES, J. W. — "The Correlation between College Grades and the Army Alpha Intelligence Tests," *Journal of Educational Psychology*, 11 : 361–367, Oct., 1920.

Bureau of Education, U. S. — *Report of the U. S. Commissioner of Education for the Year Ended*, pp. 13–15, June 30, 1920.

CALDWELL, H. H. — "Adult Tests of the Stanford Revision Applied to College Students," *Journal of Educational Psychology*, 10 : 477–488, Dec., 1919.

CALFEE, M. — "College Freshmen and Four General Intelligence Tests," *Journal of Educational Psychology*, 4 : 223–231, April, 1913.

CAROTHERS, F. E. — *Psychological Examinations of College Students*, ARCHIVES OF PSYCHOLOGY, 46; 82 pp., Dec., 1921.

CHAMBERS, O. R. — "Character Trait Tests and the Prognosis of College Achievement," *Journal of Abnormal and Social Psychology*, 20 : 303–311, Oct., 1925.

— "Columbia University Psychological Tests," *School and Society*, 13 : 141–142, Feb. 5, 1922.

— "Columbia University Psychological Examinations," *School and Society*, 17 : 268–269, March 10, 1923.

Columbia College. — *Report on the Use of Intelligence Examinations*, Columbia University, New York City; 27 pp., 1922.

COLVIN, S. S. — "Psychological Tests at Brown University," *School and Society*, 10 : 27–30, July 5, 1919.

— "The Purposes and Methods of Psychological Tests in Schools and Colleges," *Education*, 40 : 404–416, March, 1920.

— "The Validity of Psychological Tests for College Entrance," *Educational Review*, 60 : 7–17, June, 1920.

— "Educational Guidance and Tests in College," *Journal of Applied Psychology*, 5 : 32–38, March, 1921.

COLVIN, S. S., and MCPHAIL, A. H. — "The Value of Psychological Tests at Brown University," *School and Society*, 16 : 113–122, July 29, 1922.

DAVIS, H. — "Army Alpha and Student's Grades, Illustrating the Value of the Regression Equation," *School and Society*, 14 : 223–227, Sept. 24, 1921.

DECAMP, J. E. — "Studies in Mental Tests," *School and Society*, 14 : 254–258, Oct. 1, 1921.

GOWEN, J. W., and GOOCH, M. — "The Mental Attainments of College Students in Relation to Previous Training," *Journal of Educational Psychology*, 16 : 547–568, Nov., 1925.

HAGGERTY, M. E. — "Tests for Applicants for Admission to University of Minnesota Medical School," *Journal of Educational Psychology*, 9 : 278–286, May, 1918.

HOKE, E. — "Intelligence Tests and College Success," *Journal of Educational Research*, 6 : 2, Sept., 1922.

HUGHES, W. H. — "General Principles and Results of Rating Trait Characteristics," *Journal of Educational Method*, 4 : 421–431, June, 1925.

— "A Rating Scale for Individual Capacities, Attitudes, and Interests," *Journal of Educational Method*, 3 : 56–65, Oct., 1923.

— "Personality Traits and College Success of High School Graduates," *California Quarterly of Secondary Education*, 1 : 225–235, Jan., 1926.

HUNTER, H. T. — "Intelligence Tests at Southern Methodist University," *School and Society*, 10 : 437–440, Oct. 11, 1919.

Illinois, University of — *Progress Report Made to the Faculty by the Committee on Aptitude Tests*, College of Engineering, W. H. Raynor, Chairman. March, 1922.

JOHNSTON, J. B. — "Predicting Success or Failure in College at Time of Entrance," *School and Society*, 19 : 772–776, June 8, 1924; 20 : 29–32, July 5, 1924.

— "Tests for Ability before College Entrance," *School and Society*, 15 : 345–353, April 1, 1922.

JORDAN, A. M. — "Correlations of Four Intelligence Tests with the Grades of Students and with Each Other," *Psychological Bulletin*, 19 : 93–94, Feb., 1922.

— "Correlations of Four Intelligence Tests with Grades." *Ref.* Ch. 19.

— "The Validation of Intelligence Tests." *Ref.* Ch. 19.

KITSON, H. D. — *The Scientific Study of the College Student*, PSYCHOLOGICAL MONOGRAPHS, Vol. 23, no. 98; 81 pp., 1917.

LINCOLN, E. A. — "Relative Standing of Pupils in High School, in Early College and College Entrance Examinations," *School and Society*, 5 : 417–420, April 7, 1917.

MACKOWN, H. C. — *The Trend of College Entrance Requirements*, Bureau of Education, Washington, D. C.; 1924, no. 35, 172 pp.

MACPHAIL, A. H. — *Intelligence of College Students*. Warwick and York, Baltimore; 176 pp., 1924. (Contains bibliography of 294 references.)

MAY, M. A. — "Predicting Academic Success," *Journal of Educational Psychology*, 14 : 429–440, Oct., 1923.

PERRIN, F. A. C. — "Freshmen Psychological Examination in the College of Arts and Sciences at the University of Texas," *University of Texas Bulletin, 2445*, Dec. 1, 1924.

PRESSEY, S. L. — "A Group Scale for Investigating the Emotions," *Journal of Abnormal and Social Psychology*, 16 : 55–64, April, 1921.

PROCTOR, W. M. — "The High School's Interest in Methods of Selecting Students for College Admission," *School and Society*, 22 : 441–448, Oct. 19, 1925.

ROBERTS, A. C. — "Objective Measures of Intelligence Relative to High School and College Administration," *Educational Administration and Supervision*, 8 : 531–540, Dec., 1922.

ROGERS, A. L. — "The Use of Psychological Tests in the Administration

of Colleges of Liberal Arts for Women," in *Twenty-First Yearbook National Society for the Study of Education*, 1922.

ROGERS, D. C. — "Intelligence Examinations and College Entrance," *Smith Alumnæ Quarterly*, 13 : 1, Nov., 1921.

ROWLAND, E., and LOWDEN, G. — "Report of Psychological Tests at Reed College," *Journal of Experimental Psychology*, 1 : 211–217, June, 1916.

SCATES, D. E. — "A Study of High School and First-Year University Grades," *School Review*, 32 : 182–192, Dec., 1924.

SEASHORE, C. E. — "College Placement Examinations," *School and Society*, 20 : 575–577, Nov. 8, 1924.

— "Notes on Experience in Use of Mental Tests in College," *Summary of Reports made by Different Representatives in Conference at Princeton, New Jersey*, National Research Council. Mimeographed, 13 pp., Dec. 30, 1921.

— "Progressive Adjustment *versus* Entrance Elimination in a State University," *Proceedings of the National Association of State Universities*, Chapel Hill, North Carolina, pp. 15–23, 1922.

SMITH, F. C. — "The Relation between College Failures and Mental Tests," *School and Society*, 16 : 444–446, Oct. 14, 1922.

— *A Rational Basis for Determining the Fitness for College Entrance*, UNIVERSITY OF IOWA STUDIES IN EDUCATION, VOL. i, NO. 3, 1910.

STODDARD, G. D. — *Iowa Placement Examinations*, UNIVERSITY OF IOWA STUDIES IN EDUCATION, VOL. iii, NO. 2, 1925.

STONE, C. L. — "Intelligence and Scholarship," *The Dartmouth Alumni Magazine*, March, 1920.

— "The Significance of Alpha in Colleges," *Journal of Educational Psychology*, 13 : 218, May, 1922.

SYMONDS, P. M. — "The Significance of Intelligence Tests in the University of Hawaii," *School and Society*, 20 : 601–606, Nov. 8, 1924.

TERMAN, L. M. — "Intelligence Tests in Colleges and Universities," *School and Society*, 13 : 481–494, April 23, 1921.

THORNDIKE, E. L. — "The Future of the College Entrance Examining Board," *Educational Review*, 31 : 470–479, May, 1906.

— "Entrance Exams and College Grades," *Science*, 23 : 839–845, June, 1906.

— "Measurement in Education." NATIONAL SOCIETY FOR THE STUDY OF EDUCATION, TWENTY-FIRST YEARBOOK. Part I. *Intelligence Tests and Their Use*, 1922.

THORNDIKE, E. L. — "New Psychological Tests: Are They a Substitute for Content Examinations?" *Educational Review*, 59 : 97–104, Feb., 1920.

— "Intelligence Examinations for College Entrance," *Journal of Educational Research*, 1 : 329–337, May, 1920.

THURSTONE, L. L. — "Mental Tests for College Entrance," *Journal of Educational Research*, 10 : 129–142, March, 1919.

— "Mental Tests for College Entrance," *Proceedings of Educational Association of Western Pennsylvania;* pp. 16 f., 1918.

— "Psychological Tests for College Freshmen," *Educational Record*, 6 : 69–83, April, 1925; 6 : 282–294, Oct., 1925.

— "Intelligence Tests for Engineering Students," *Engineering Education*, 13 : 263–318, Jan., 1923.

TOOPS, H. A. — "The Status of University Intelligence Tests in 1923–1924," *Journal of Educational Psychology*, 17 : 23–36; 110–124, Jan., Feb., 1926.

WALCOTT, G. D. — "Mental Testing at Hamlin University," *School and Society*, 10 : 57–60, July 12, 1919.

WHIPPLE, G. M. — "Intelligence Tests in Colleges and Universities." *Ref.* Ch. 4.

WHITMAN, A. D. — *The Value of College Entrance Examination Board Examinations on Prediction of College Success.* Bureau of Publications, Teachers College, Columbia University, 27 pp., 1926.

WOOD, B. D. — *Measurement in Higher Education. Ref.* Ch. 2.

"The College Curriculum and Vocational Guidance," *School and Society*, 21 : 508–514, April 25, 1925.

CHAPTER XXI

THE USE OF TESTS IN GUIDANCE

For over a decade many persons have written and talked about the possibilities of advising boys and girls as to their educational and vocational careers. Adequate guidance is especially needed in the life of to-day with its specialization of activity and opportunity. The latest U. S. Bureau of Education [1] figures indicate that some 201,000 pupils dropped out of school at the end of the 8th grade in 1920; 278,000 pupils dropped out of school at the end of the 9th grade in 1921; 201,000 at the end of the 10th grade in 1922; and 81,000 at the end of the 11th grade in 1923. Probably only a very small fraction of those who dropped out at the end of the 8th grade had any adequate preparation for a vocation or any definite idea as to what they were going to do. Some drifted into local factories and stores as errand boys, or delivery boys, some stayed at home to work on the farm, some merely loafed about until something turned up. Nearly all of them were in doubt as to whether to continue on into high school. Some, whose parents wished them to continue, dropped out because school had nothing to interest them or to promise them. Some wished to continue in school, but dropped out because their parents wanted them to go to work. Practically all were much in need of guidance as to what they should do.

There are three factors which make guidance a need peculiar to present-day life and one which becomes more pressing as these conditions become more pronounced. In the first place, there is the enlarged curriculum, which is a feature of secondary edu-

[1] "Statistics of State School Systems 1923–1924," *U. S. Bureau of Education Bulletin no. 42, 1925.*

cation in city systems. At the end of the elementary school a cosmopolitan high school offers several alternatives. A recent study by Counts[1] of the curriculum situation in secondary education reveals that the curriculum offerings in American cities are many and diverse. Pueblo offers two curricula, whereas Los Angeles offers eighteen curricula. These curricula have many different names, as the following table illustrates.

TABLE 117

CURRICULA OFFERED IN HIGH SCHOOL IN 15 AMERICAN CITIES ADAPTED (FROM COUNTS)

College preparatory	11	Electrical industry	2
Commercial	9	Architecture	1
General	8	Chemistry	1
Home economics	6	Public Health	1
Science	6	Academic	1
Bookkeeping and accounting	6	Literature and arts	1
Teacher-training	5	Music and art	1
Agriculture	4	Business administration	1
Classical	4	Dressmaking	1
Business	4	Home nursing	1
Stenography	4	Commercial food service	1
Engineering	3	Art	1
Secretarial service	3	Literary	1
Mechanic arts	3	Social service	1
Printing	3	Salesmanship	1
Technical	2	Building industry	1
English	2	Modern language	1
Medicine	2	Photography	1
Clerical service	2	Household arts	1
Industrial arts	2	Machine work	1
Fine Arts	2	Pattern and cabinet making	1
Music	2	Carpentry	1
Drafting	2	Retail distribution	1
Automobile industry	2	Manual training	1

[1] COUNTS, G. S. — *The Senior High School Curriculum*, SUPPLEMENTARY EDUCATIONAL MONOGRAPHS PUBLISHED IN CONJUNCTION WITH THE SCHOOL REVIEW AND THE ELEMENTARY SCHOOL JOURNAL, NO. 29, Feb., 1926.

Some of these probably represent the same curriculum but under a different name. Even so, there are many distinct lines of activity. The main channels seem to be preparatory, home economics, commercial, manual training, and teacher-training. The varied program of almost every city high school is too confusing to be unraveled by a pupil entering high school. He needs advice and guidance.

A second factor which makes guidance a necessity is the heterogeneous character of the high school population. Thirty or forty years ago the high school's sole function was to prepare for college except for a little commercial work done in the more progressive schools. If a pupil did not wish to prepare for college, he took the same course without Latin, called the general course. But to-day pupils of many interests and needs are entering high school. Some will stay but a year or a fraction of a year — others will go on much further. Where all who entered high school had a common purpose guidance was unnecessary. To-day the school offers a more extensive curriculum and there is need of guidance for the more heterogeneous group of students.

A third factor which makes guidance a necessity is the specialization of occupations. This well-known phenomenon does not require further treatment here.

One phase of guidance is pupil enlightenment as to school offerings and vocational possibilities. The modern large city high school is so complex that probably no teacher knows all its ramifications. The school offerings can be described in catalogues, but these catalogues need interpretation for pupils and parents. A specialist is needed to hold group and individual conferences for giving the explanations and advice concerning the school offerings. This must be done systematically. The matter of vocational enlightenment is more difficult as it depends on a knowledge of the character of the work in different occupations, the remuneration, the training necessary, the opportunities for advancement,

and the local opportunities. This data must be constantly assembled, and the services of a bureau or special worker is needed for such work. Then there is the matter of disseminating this information among the student body of a school in a way that is easily understood.

A second phase of guidance is the trial course. The trial course is usually a short unit of work or activity which a pupil may take to try himself out. Such courses, if wisely administered, enable a pupil to direct his course by trying out various activities. There is a danger here that superficial factors may influence choice and may quicken or deaden interest to pursue an activity further. These superficial factors, such as liking for the teacher or early success in the subject, may obscure the more permanent abilities and interests that well-constructed tests may reveal.

A third phase of guidance is organization of the school to care for individual differences. Ability grouping, discussed in a later chapter, is in a sense a form of guidance in that it enables a pupil to make a choice of his future activity. There are other forms of school organization that may properly be called a form of guidance, such as special classes for repeaters, supervised study, special Saturday and summer coaching classes. Certain New York City schools dismiss all except failures during the last week of the term and give these failures intensive work of several hours a day.

It is in the use of tests, however, that the greatest hope for scientific guidance lies. If, by means of a test, one can obtain a sample performance that is symptomatic of ability which is predictive, then one has a ground for offering advice on the basis of probabilities of future success or failure.

Thorndike [1] was among the first to perceive the possibility of using tests for guidance. As early as 1913, in an address given at a Teachers College Alumni meeting, he outlined the possibility of

[1] Thorndike, E. L. — "Address at Teachers College Alumni Meeting," *Teachers College Bulletin*, 15: 5–15, Mar., 1913.

giving vocational and educational guidance on the basis of exact and comprehensive investigations of occupations and measurement of human capacities. Even at that time he outlined what is to-day a reality in the form of intelligence tests. Kelley, in his *Educational Guidance*, published in 1914, carried through an investigation showing the possibilities of using tests in guidance. This piece of work used the technique of partial and multiple correlation and the regression equation and has stood as a model for all later work. In the testing with Army tests during the war, data assembled concerning the intelligence level of different occupational groups were heralded as a great boon to the possibility of guidance through the use of tests. These data have not yielded the results expected, partly because of a misconception concerning the function of tests in guidance. Nothing since has been accomplished of significance except the work of Toops [1] and its follow-up by the Institute of Educational Research under the guidance of Thorndike.

Part of the failure to make tests function in guidance is that the theorists and school administrators have not come together on the problem. The former have been content to elaborate fine-spun techniques which may be used many years from now when a multitude of tests are available and people have been so converted to their use as to appropriate much money and the service of many experts for their administration. On the other hand, school administrators with a definite lack of imagination and a reluctance to experiment have tried to make general intelligence tests do the whole trick of guidance. The rest of this chapter will have served its purpose if it can interpret the work of theorists so that those who are in a position to and are willing to experiment may have suggestions which have the possibility of yielding some immediate success and progress.

First, what is the position of the theorist?

[1] *Cf. Tests for Vocational Guidance of Children Thirteen to Sixteen*, 1923.

The advantage of a battery of tests for purposes of vocational guidance, for example, would be enormous. . . . But even assuming an unlimited supply of separate teams of tests of the type now generally aimed at — the tests of each properly weighted by means of a regression equation for its particular vocation — the problem of vocational guidance would still be far from solved because of the expense involved in such a system. If each battery should require one or two hours for giving, scoring, etc., and the multiplicity of vocations should be reduced to 40 or 50 type occupations, it would require from 50 to 100 hours of labor to discover the two or three most promising occupations for each individual! But if, instead of a multiplicity of teams of tests, a single approximately universal battery of tests were available which, through the perfection of group testing by means of self-recording duplicatable apparatus and other devices, groups of 25 or 50 subjects should be tested at once, vocational guidance might easily be economical enough to become universal. By this latter method there would be merely a multiplication of regression equations, each equation weighting the *same* team of tests in a different way according to the peculiar mental requirements for success in each particular vocation. Thus but one set of tests would need to be administered and scored. And lastly there might be an automatic computing apparatus built on an extension of the principle of modern statistical machines, which would solve the regression equations. With the test data from a given subject punched down on an appropriate form, this might be fed into the machine and presently emerge with the whole series of predictions recorded in the units of some convenient uniform scale that would permit of instant comparison. This, of course, is all highly Utopian. But without some such revolution in method as suggested above, it is difficult to see how scientific vocational guidance can ever become a practical reality for the masses of the people.

Passing to the immediately practical, there seems little doubt that the near future will see a tendency towards semi-universal teams of tests — single batteries of tests which will apply to a limited variety of occupations of a given type, there being separate regression equations based on the same tests, for the various occupations of the particular group. And along with this ought also to come a specific concentration upon the psychology of the independencies among tests, with consequent change in the point of view in devising new tests and in the technique of choosing the individual tests to make up teams or batteries for purposes of vocational prognosis.[1]

[1] Hull, C. L., and Limp, C. E. — "The Differentiation of the Aptitudes of an

This is an admittedly Utopian program, yet Hull has since solved what seemed to him the key point in the program by actually constructing an automatic correlation machine. The problem is, however, far from being solved, for even with the aid of his machine Hull has not produced a universal set of tests with their regression equations to predict an individual's aptitude in 40 or 50 type occupations.

On the other hand, school administrators have for several years been working with general intelligence tests as their only tool for individual diagnosis. After the war much interest was taken in the tables of the intelligence levels of occupations which appeared. The one which follows was prepared by Fryer,[1] who utilized the Army data but corrected some of the discrepancies in these data from research carried on in the Central Branch of the Brooklyn Y. M. C. A.

Concerning this list Fryer says,

> The mean for the occupation is presented as the "score average," and the "score range," indicating the range of intelligence within which can be expected success in the occupation, secures its limits (usually, but not always) from the first and third quartiles. The scores are so presented as to indicate that in all probability an individual must have an intelligence rating within the "score range" for achievement in the occupation, with the further probability that he should be above the "score average" to be sure of sufficient intellectual capacity for the occupation.

The range in intelligence is so great for any one occupation that it is impossible to give definite occupational advice. Suppose that a boy makes a score of 96 on the Army Alpha. What advice concerning vocational choice shall he be given? There are twenty-three occupations in the above list whose "score range"

Individual by Means of Test Batteries," *Journal of Educational Psychology*, 16 : 73–88, Feb., 1925.

[1] FRYER, D. — "Occupational-Intelligence Standards," *School and Society*, 16 : 273–276, Sept. 2, 1922.

TABLE 118

OCCUPATIONAL INTELLIGENCE STANDARDS BASED ON ARMY ALPHA INTELLIGENCE TESTS (FROM FRYER)

Score Average	Score Range	Occupation	Score Average	Score Range	Occupation
161	110–183	Engineer	75	56–105	Graphotype Operator
152	124–185	Clergyman	74	53–91	Engineer (Locomotive)
137	103–155	Accountant	72	54–99	Farrier
127	107–164	Physician	70	46–95	Telephone Operator
122	97–148	Teacher	70	44–94	Stock Checker
119	94–139	Chemist	69	49–93	Carpenter (Ship)
114	84–139	Draftsman	69	48–94	Handyman (Gen'l Mechanic)
111	99–163	Y. M. C. A. Secretary	69	46–90	Policeman and Detective
110	80–128	Dentist	68	51–97	Auto Assembler
109	81–137	Executive (Minor)	68	47–89	Engineer (Marine)
103	73–124	Stenographer and Typist	68	42–86	Riveter (Hand)
101	77–127	Bookkeeper	67	50–92	Toolmaker
99	78–126	Nurse	66	45–92	Auto Engine Mechanic
96	74–121	Clerk (Office)	66	45–91	Laundryman
91	69–115	Clerk (Railroad)	66	49–86	Gunsmith
86	59–107	Photographer	66	44–88	Plumber
85	57–110	Telegrapher and Radio Operator	66	44–88	Pipe Fitter
83	64–106	Conductor (Railroad)	65	44–91	Lathe Hand (Production)
82	57–108	Musician (Band)	65	43–91	Auto Mechanic (General)
81	59–106	Artist (Sign Letterer)	65	43–91	Auto Chauffeur
81	60–106	Clerk (Postal)	65	42–89	Tailor
81	57–109	Electrician	65	44–88	Carpenter (Bridge)
80	62–114	Foreman (Construction)	64	43–83	Lineman
80	56–105	Clerk (Stock)	63	40–89	Machinist (General)
78	54–102	Clerk (Receiving and Shipping)	63	46–88	Motor Cyclist
78	61–106	Druggist			
77	59–107	Foreman (Factory)			

TABLE 118 (*Continued*)

Score Average	Score Range	Occupation
63	41–86	Brakeman (Railroad)
62	31–94	Actor (Vaudeville)
61	40–85	Butcher
61	44–84	Fireman (Locomotive)
61	39–82	Blacksmith (General)
60	38–94	Shop Mechanic (Railroad)
60	36–93	Printer
60	40–84	Carpenter (General)
59	40–87	Baker
59	39–83	Mine Drill Runner
59	38–81	Painter
58	37–85	Concrete Worker
58	40–83	Farmer
58	37–83	Auto Truck Chauffeur
58	37–82	Bricklayer
57	41–81	Caterer
57	39–71	Horse Trainer
56	38–76	Cobbler
55	35–81	Engineer (Stationary)
55	34–78	Barber
55	35–77	Horse Hostler
52	38–96	Sales Clerk
52	33–74	Horse Shoer
51	31–79	Storekeeper (Factory)
51	26–77	Aeroplane Worker
51	31–74	Boiler Maker
50	33–75	Rigger
50	30–72	Teamster
49	40–71	Miner (General)
48	21–89	Station Agent (General)
40	19–67	Hospital Attendant
40	19–60	Mason
35	18–62	Lumberman
35	19–57	Shoemaker
32	16–59	Sailor
31	20–62	Structural Steel Worker
31	19–60	Canvas Maker
30	16–41	Leather Worker
27	19–63	Fireman (Stationary)
27	17–57	Cook
26	18–60	Textile Worker
22	16–46	Sheet Metal Worker
21	13–47	Laborer (Construction)
20	15–51	Fisherman

includes a score of 96 from *farrier* to *chemist*. The only advice that can be given is of a negative sort. One would advise a boy with a score of 96 not to plan to be an engineer as he would at best be only a poor engineer and would have much competition for success by those with more ability to start with.

One would also advise him to aim at something higher than *fisherman* or *miner*, if such advice was necessary.

Definite educational advice cannot be given from the data on an intelligence test. One cannot advise a boy on the basis of an intelligence test whether or not to drop or continue school or what course of study to take. The most that can be done is to give advice of a negative sort. Sufficient evidence has been gathered to permit Terman [1] to make such statements as: "In this high school, at least (Palo Alto), the pupil with I Q below 90 is practically certain to fail in such studies as algebra and Latin." "Below 90 I Q graduation (from high school) is by no means likely." "The chances are remote that a child testing much below 90 will ever be able to satisfy the requirements for college graduation. Children who test below 100 should ordinarily not be encouraged to look forward to entrance into law, medicine, the ministry, engineering, teaching, or any other profession which demands a high degree of ability in abstract or conceptual thinking." "Anything above 85 I Q in the case of a barber probably represents so much waste." Miss Cobb [2] says, "Probably in 90 cases out of 100, it is unwise to guide the average or less intelligent than average child into the present academic high school. Unless his I Q is over 100, or his mental age definitely over 14, he should be encouraged to try some other type of training." Such advice, thin and meagre as it is, is of immense value in preventing *serious* mistakes of educational and vocational choice which are only too common.

Similar comments might be made concerning the value of prognosis tests for guidance when such tests have been constructed. Prognosis tests will predict with some precision (r = .70 to .80)

[1] Terman, L. M. — *The Intelligence of School Children*, pp. 86, 90, 281, 282, 288; 1919.

[2] Cobb, M. V. — "The Limits Set to Educational Achievement by Limited Intelligence," *Journal of Educational Psychology*, 13: 449–464, 546–555, Nov., Dec., 1922.

probable success with a given subject, but that is not a sufficient basis for choosing that subject. Suppose that a set of prognosis tests revealed to a boy that he would be successful with Latin, French, German, algebra, biology, chemistry, physics, dramatics, bookkeeping, typewriting, stenography, etc, etc. Such information does not tell him which of these studies he should take or which line he should follow. Or if the tests tell another boy that he will be a failure in all of these studies, this boy is not given advice as to what he should do.

We are forced to the conclusion that a single test does not furnish information which may be used for guidance. One is frequently asked to recommend a test "to tell me whether I should take this subject." A *single* test will not give such information for any subject.

The information needed as a basis for guidance is of the sort "You will be relatively more successful in this subject than in that" or "You will be relatively more successful in this vocation than in that." This sort of information requires at least two tests. As a general principle it may be said that *information concerning abilities as a basis for guidance requires the use of at least two tests.* Guidance information should always be comparative.

Elementary education in America is uniform and integrative. The pupil has one choice — the choice is made by those who determine the curriculum. The pupil's first opportunity for choice comes somewhere in the junior high school. The first choice, and in many respects the most important one, in a child's whole educational career comes when he finishes his elementary education and makes the choice that leads into the academic, commercial, mechanic arts, or household arts courses. Tests are needed which will inform a pupil in which of these courses his abilities are apt to be most rewarded by success. Probably tests will never be available to predict ability in separate occupations. We must begin at least with the more general divisions of

activity. In order to work out a usable system of tests for educational guidance to perform this first differentiation of pupils, two things must be accomplished. *Tests must be found which correlate most highly with ability in school in academic, commercial, mechanic arts, and household arts courses.* So far as is known, general intelligence tests correlate most highly with academic success. (The data are summarized in chapter XIX. Tests for predicting success in commercial or mechanic arts courses were discussed in chapter XVIII.) The second requirement is that *accurate norms of performance be available for all tests which have been obtained on the same pupils.* No tests are really of use for the purposes of guidance until norms are available. When such tests have been assembled, the relative abilities in the different fields may be displayed graphically, as indicated in the following figure.

Figure 13 tells at a glance relative abilities in these four fields.

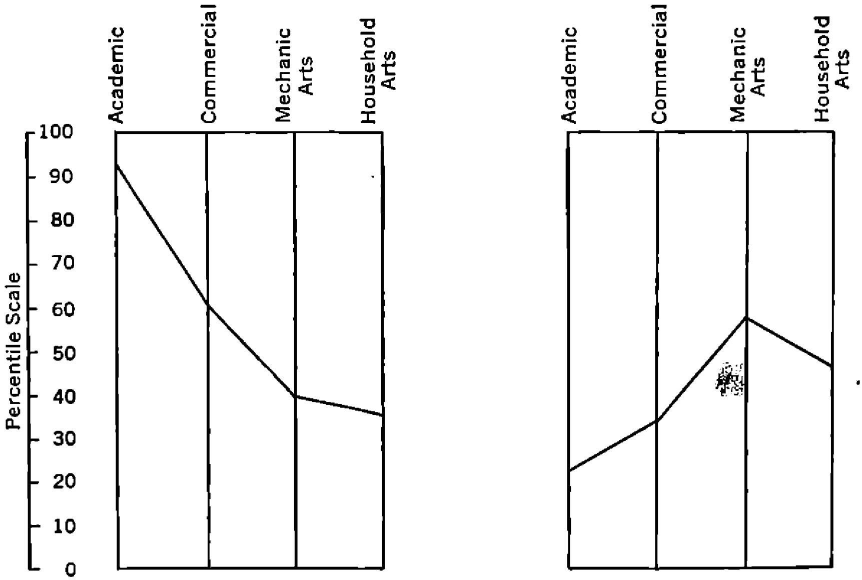


FIGURE 13. — Profile graph suggested for the guidance of pupils in high school

The scale at the left is intended to represent the percentile scale of ability of all pupils who are in high school, of whatever curriculum. The vertical lines represent the abilities on tests designed to correlate highly with success in each of the curricula: academic, commercial, mechanic arts, and household arts. Lines showing the possibilities of variation have been drawn connecting hypothetical points which might represent the standing of two pupils. Because of the low correlations between abstract and mechanical abilities this profile would probably show extreme variations up and down. Stenquist has tellingly insisted that we are wrong in estimating ability by means of abstract tests alone, because the low correlations between abstract and mechanical ability makes it possible that one may be high in mechanical ability and low in abstract ability. All possible combinations would be possible in the above graphs and different advice would be given for each of the combinations. Such a set of tests that could be used for guidance is distinctly a work for the future, but an excellent groundwork has been laid in the work on abstract, clerical, and mechanical ability tests.

The scheme just outlined is offered as a bridge between the theorists and the practical work in the schools. It at least offers a program for research which promises immediate success and a more immediate set of tools for guidance than a program involving 40 or more tests. It must be admitted that part of the significance of the data is not made use of in this graphic method. The regression equation gets the most out of whatever tests are used and should be employed wherever possible. But in planning a research program in the construction of tests for guidance let us lay plans that may be completed in this generation.

Interest is also a factor in the prediction of success. Kelley,[1] in his original study, tried out interest questionnaires and found that they had some correlation with later school success.

[1] *Cf. Educational Guidance*, pp. 40–62.

Freyd [1] worked out a set of questions which were found to differentiate sharply between the mechanical minded (engineering students) and the social minded (insurance salesmen). More recently Cowdery,[2] Strong, and others at Stanford have devised a simple form of questionnaire for college students concerning occupational interests and other interests which differentiates between those who will be successful doctors, lawyers, and engineers.

The use of interest questionnaires to determine differential aptitudes in the junior high school is a field of research that has not been exploited and needs considerable investigation. Interest questionnaires may serve as a valuable complement to ability tests. It is probable that interest questionnaires are more significant at higher ages when interests have become more pronounced and definite than in the junior high school age when they are more or less fluid and indefinite. However, this is a matter for research to decide.

It should be added that there are other factors besides abilities and interest that should be influential in determining educational choice. Should one take French or German? Here ability or interest probably is not a factor at all in the majority of cases. In many cases such a matter is to be answered only by such considerations as what one's friends take, which teacher is the more popular, which is considered more socially desirable, or whether the language is that of a country which was friend or foe in the recent war. In college I spent many painful hours trying to decide whether to take a course in "Symbolic Logic" or "The Foundations of Mathematics." I chose the former. Such decisions are not very important. The important decisions are those in which differentiation of ability really counts.

[1] Freyd, M. — *The Personalities of the Socially and Mechanically Inclined*, Psychological Monographs, no. 151, 1924.

[2] Cowdery, K. M. — "Measurement of Professional Attitudes," *Journal of Personnel Research*, 5: 131–141, Aug., 1926.

Résumé. At the present time there are no tests except general intelligence tests that may be used for the purposes of guidance. The advice that one can give from an intelligence test is extremely vague and general and on the whole negative. But this advice is of the highest importance. One can advise a boy on the basis of his I Q that he should or should not plan to take the academic course in high school; or should or should not look forward to college. One can advise pupils with I Q's below 110 not to take algebra or Latin. From the information yielded by a general intelligence test one cannot tell a pupil what course he should follow, or in which of two alternative courses or subjects he will be more successful.

The information needed in order to give adequate guidance is that which tells in which of two subjects or courses a pupil is likely to have more success. This advice requires at least two tests. These tests should correlate as highly as possible with the two subjects or courses between which choice is to be made. Norms are needed to interpret the results of the tests.

The most pressing research problem in guidance is the construction of tests which will predict relative success in the work of the academic, commercial, mechanic arts, and household arts courses. A basis has been laid for this in the work of Toops and others in determining the predictive value of abstract, clerical, and mechanical ability tests, but more must be done in trying them out on standard groups before they will be available for the purposes of guidance.

REFERENCES

ALLEN, R. D. — "Educational Guidance in High Schools," *National Vocational Guidance Bulletin*, 1 : 83–86, Jan., 1923.

— "Some Neglected Problems in Classifying High-School Pupils." *Ref.* Ch. 15.

BAKER, H. G. — "Objective Measurement in Educational and Vocational

Guidance" in *Twenty-Third Yearbook, National Society for the Study of Education, Part II*, pp. 151–172, 1924.

BREWER, J. M. — "Guidance in the High School with Special Reference to College Entrance," *School Review*, 29 : 434–443, June, 1921.

BREWER, J. M., and others — *Mental Measurement in Educational and Vocational Guidance*, HARVARD BULLETIN ON EDUCATION, NO. 10. Harvard University Press, Cambridge, 1924.

COBB, M. V. — *Op. cit.*, Ch. 13.

COUNTS, G. S. — *The Senior High School Curriculum*, SUPPLEMENTARY EDUCATIONAL MONOGRAPHS, NO. 29, published in conjunction with the School Review and the Elementary School Journal, Feb., 1926.

COWDERY, K. M. — "Measurement of Professional Attitudes. Differences between Lawyers, Physicians, and Engineers," *Journal of Personnel Research*, 5 : 131–141, Aug., 1926.

DICKSON, V. E. — "The Need of a Counseling Program in Secondary Schools," *Journal of Educational Research*, 11 : 12–16, Jan., 1925.

FORBES, A. W. — "Occupation-Intelligence Standards: A Reply to Mr. Fryer," *School and Society*, 16 : 414–416, Oct. 7, 1922.

FREYD, M. — *The Personality of the Socially and Mechanically Inclined*, PSYCHOLOGICAL MONOGRAPHS, NO. 151, 1924.

FRYER, D. — "Occupational-Intelligence Standards," *School and Society*, 16 : 273–276, Sept. 2, 1922.

HULL, C. L. — "An Automatic Machine for Making Multiple Aptitude Forecasts," *Journal of Educational Psychology*, 16 : 593–598, Dec., 1925.

— "The Joint Yield of Teams of Tests," *Journal of Educational Psychology*, 14 : 396–406, Oct., 1923.

HULL, C. L., and LIMP, C. E. — "The Differentiation of the Aptitudes of an Individual by Means of Test Batteries." *Ref.* Ch. 17.

KELLEY, T. L. — *Educational Guidance. Ref.* Ch. 17.

KITSON, H. D. — "Vocational Guidance and the Theory of Probability," *School Review*, 28 : 143–150, Feb., 1920.

MADSEN, I. N. — "The Contribution of Intelligence Tests to Educational Guidance in High School," *School Review*, 30 : 692–701, Nov., 1922.

MARTENS, E. — "A High School Counseling System in Operation," *Journal of Educational Research*, 11 : 17–24, Jan., 1925.

PROCTOR, W. M. — *Educational and Vocational Guidance*. Houghton Mifflin Company, Boston; 352 pp., 1925.

— *Psychological Tests and Guidance of High School Pupils. Ref.* Ch. 19.

TERMAN, L. M. — *The Intelligence of School Children. Ref.* Ch. 15.

THORNDIKE, E. L. — "Address at Teachers College Alumni Meeting," *Teachers College Bulletin*, 15 : 5–15, March, 1913.

— "The Relation of Interests to Abilities," *School and Society*, 5 : 178–179, Feb. 10, 1917.

TOOPS, H. A. — *Tests for Vocational Guidance of Children.* *Ref.* Ch. 18.

U. S. Bureau of Education, *Bulletin 1925, no. 42.* "Statistics of State School Systems," 1923–1924.

CHAPTER XXII

PROMOTION

All teachers are bothered by the necessity of having to make decisions concerning the passing or failing of their pupils. The difficulty comes partly from working individually in the determination of marks, rather than as a faculty. No teacher in a school can stand alone and independent in the assignment of marks. If the point of view were changed so that the problem became one of *pupil placement* rather than of *passing* or *failing*, much of the confusion regarding the matter would be clarified, even though the problem would not be simplified. This chapter and the following one are devoted to a discussion of pupil placement — in this chapter the matter of promotion is treated and in the next that of sectioning.

There are two fundamental principles of pupil placement. For such subjects as algebra, geometry, history, civics, the sciences, and, in general, subjects which are more or less independent and can be taken as isolated units, the principle of placement is that *no pupil should be admitted to a course who does not have an even chance of achieving a certain standard in the course.* Prognosis tests are needed for all subjects where this principle applies. At the present time there are few really satisfactory prognosis tests in any high school subject. As now administered, admission to a course is determined by the satisfactory accomplishment of two requirements: (1) that the work of the previous term shall have been passed, (2) that all prerequisites shall have been taken and passed. Both of these may be criticized. The passing of any term's work is highly subjective. The demand of prerequi-

sites has probably been very much overworked. Investigations are needed to show that pupils may take geometry, physics, or chemistry without having had algebra. Of course it is granted that algebra is a tool in those subjects, but if it can be shown that achievement is satisfactory in geometry or the sciences without first having had algebra, the requirement of algebra as a prerequisite has been overstressed. A prognosis test takes the place of both of the present requirements. It measures achievement in material that predicts success in a subject; and a prognosis test will contain the equivalent of the prerequisites if statistical analysis shows such prerequisites to be necessary. Since there are few prognosis tests at the present time, it is useless to show in detail how they may be used. The main thing is to get the prognosis tests constructed — their use is simple. It consists in giving the tests prior to the promotion of a class and admitting to the next class on the basis of standards which must be set for the purpose of promotion.

Passing or failing at the end of the course should be determined by means of standardized tests, using as the goal of prediction the same standards used in the prognosis tests.

The second principle of pupil placement (to be used with English, foreign languages, and other subjects continuous over several terms) *is that a pupil should be placed in a group where the material to be learned is most nearly equivalent in difficulty to the pupil's ability.* In order to apply this principle one must know something of the law of growth in a group. The nature of this law is shown by the following diagram.

Let the scale at the base be a scale of achievement as measured by some standardized test. Let the curves at various levels illustrate the frequency distribution of achievement in successive terms. Note that the mean of the curves progresses steadily to the right, indicating average progress through the various terms. Note that the variability increases. This indicates that the learn-

ing of pupils is uneven, some learning much faster than others. It is this fact of increasing variability in the group that causes the need for pupil placement. If the variability remained constant, the group would remain relatively homogeneous and could be promoted as a group — none would fail. As it is, some outstrip the others and soon reach an achievement equivalent to that of some group ahead; and some lag behind so that another group

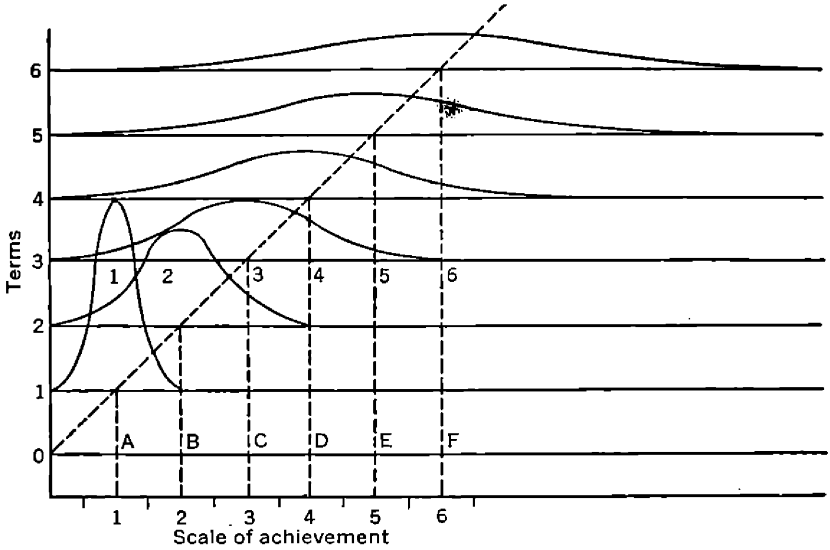


FIGURE 14. — The growth of a school group in some subject

overtakes them in achievement. Since school organization necessitates teaching in classes, some of the fast-moving pupils should eventually be skipped to the group ahead; and some should be kept back (or discouraged from proceeding further). It is this problem of determining who should be kept back, who should remain with the group, who should be skipped to the next group ahead, that constitutes the problem of pupil placement.

The vertical lines run through the mean point of each of the terms, and these mean points in achievement are lettered *A*, *B*, *C*, *D*, *E*, and *F*. Midpoints between each of these letters are desig-

nated on the lowest scale and these midpoints mark off the range of achievement that clusters around the mean of achievement in any term. These ranges are numbered according to the terms of which they are representative. If these ranges are applied to any term, say, the third, it will be noted that the pupils in the third term overlap several terms in achievement; some are of first-term achievement, some are of second-term achievement, some are of third-term achievement, and some of fourth. Insofar as the test records are true measures of achievement the evidence shows that only a fraction of the third-term class should be promoted to fourth-term work. Some should be skipped a term to fifth-term work; some should be held back to repeat the third term; and some should be actually demoted to the second term.

McCall,[1] who has done the most extensive work in connection with problems of pupil placement, has worked out the following rules for classification which I repeat here, changing the terminology slightly to accord with high school conditions.

1. *No pupil will be demoted or denied his normal promotion unless his educational achievement falls below the third quarter of the term in which it is proposed to place him.* This rule gives the slow pupil the benefit of any doubt concerning the reliability or validity of the tests. Further, it may not be wise to run the risk of discouraging by demotion or failure of promotion a stupid pupil who may be doing his best unless it is pretty certain that he really needs the work below.

2. *No pupil will be skipped over a term whose educational achievement does not exceed the first quarter of the term to which it is proposed to skip him.* If the classifier cannot afford to risk a few failures as a result of his recommendations, he had better require that the pupil exceed the median of the term to which he is promoted.

What is needed by the teacher in order to put this plan

[1] *Cf. How to Measure in Education,* p. 48.

into effect, and to *pass* or *fail* in order to insure pupil placement in accordance with the above plan? In the first place, there must be available tests which are reliable and which the teacher is satisfied adequately measure the objectives of the course. In the second place, there must be standards for the test which are agreed upon by all the teachers of the school. These standards may be norms obtained for the test as published, or norms may be obtained by any school for its own use from its own past records. With tests and standards available the problem of passing or failing becomes greatly simplified for the teacher. Passing and failing should be administered entirely for the purpose of pupil placement. Other functions of marks, such as determining credits for graduation, are only secondary. Passing or failing for the purpose of graduation again needs only to be considered from the point of view of pupil placement.

The peculiar nature of secondary education makes the problem of pupil placement more complicated than in the elementary school where a pupil within the compulsory attendance age limits *has* to be placed in *some* grade. The problem becomes purely one of placement. But in the secondary school differentiation begins to take place. Pupils may choose their courses to some extent. As the work gets more advanced, more choice is left to the pupil as to what he shall take. Even in the early years in the high school, if a pupil is dissatisfied with his placement, he may leave the course or leave school altogether. These alternatives, which must be taken into account in pupil placement, help modify the general principles of pupil placement.

In our illustration many pupils in the third term would have to repeat the work of that term if the above suggestions were to take effect. Repeaters are a troublesome problem in high school. If the school is small, they have to be placed in the oncoming classes, — if the school is large, special repeating classes may be formed. If special repeating classes are possible, it should be insisted that

the work be varied in some respect. There is nothing so deadening as repeating exactly the same work. Something should be changed. Change texts if possible, change teachers, change order of topics. The teacher of a repeating class should use every motivating device at her command.

However, from the modern point of view, it may be well to advise pupils who are unable to keep up with their class to drop the subject altogether. A repeating class almost defeats the purpose of secondary education. For secondary subjects represent a beginning of specialization, and no one should expect to specialize without a promise of success. If a boy takes algebra and cannot keep up with his class, algebra is not worth anything to him — there are too many others who can master algebra and profit by it. The same is true of the languages, the sciences, and of the *advanced* social studies. High school and college education have become fashionable and too many parents try to force their children through the course more for the prestige that accompanies the diploma than for any real benefit from the courses. It would not make any difference to such parents what their children study in school so long as they are studying the same subjects that the élite study. So a desirable alternative to requiring a pupil to repeat the work of a year is advising him to select some other course where his efforts are likely to be better rewarded with success.

These remarks will have more meaning if illustrated by data from actual school experience. Following is a table presenting distributions in first-, second-, and third-year French, as tested by the vocabulary test in Part I of the "American Council French Test."

These distributions indicate that actual school conditions are like the theoretical conditions pictured in figure 14. The average achievement in French moves steadily upward each year and the variability of each succeeding class becomes larger. Note the

TABLE 119

DISTRIBUTION ON FRENCH VOCABULARY TEST (AMERICAN COUNCIL TEST)

SCORE	FIRST YEAR FRENCH	SECOND YEAR FRENCH	THIRD YEAR FRENCH
65–69			1
60–64			
55–59			1
50–54			1
45–49			5
40–44		1	4
35–39		2	6
30–34	1	2	8
25–29	8	18	3
20–24	11	12	1
15–19	12	8	
10–14	6	1	
5–9	1		
1–4			
	39	44	30

large amount of overlapping. One pupil in second-year French is only slightly above the lowest pupil in first-year French; and one pupil in first-year French is well above the average of achievement of second-year French. These are conditions of which the teachers were unaware in the school from which these data were obtained (a large private preparatory school of high standard). If these two teachers had expressed their opinion, they would have asserted that none of the first-year French pupils could do the work of second-year French. And yet the conditions found in this school are typical of conditions in every school in the country. Wood, in extensive studies in modern languages in New York City Junior High Schools, found these conditions wherever he investigated. He also found wide variations in achievement in supposedly second-year French and third-year

French classes, yet all were being taught approximately the same material.

It would be encouraging to see a high school language department that is courageous enough to break traditional lines of the pass-fail placement and try the placement of pupils on the basis of their recorded achievement. This would mean that the continuity of the course would have to be broken for some pupils in grammar. There are two ways of handling this. One is to ignore the lack of continuity until it has been proved that the formal work in grammar is of as much importance as many language teachers consider it; the other is that pupils, if granted the privilege of faster-than-average progress, might be required to do extra work in grammar and pass a special examination before being permitted to skip work.

The distributions presented in table 119 represent the testing of vocabulary alone. The American Council Test also contains tests of grammar knowledge and reading ability. It would be unwise to determine pupil placement on the basis of one of these tests alone, but each of the three measures French achievement from important yet different aspects and the three taken together would seem to provide a satisfactory basis for placement.

Passing and failing can only have an accurate and definite meaning when given for the purpose of pupil placement and by means of standardized tests. It would seem that these terms should be reserved for use in connection with the transition from one class to another, or at the completion of a term's work; and after the taking of a standardized test which can only be given two or three times a year. The use of passing and failing in connection with daily work should be abandoned. Although pupils require frequent measures of their achievement, these measures do not need to be always interpreted in terms of passing and failing. I know of no way of satisfactorily interpreting a traditional test, or

an informal test of the new type, so as to yield information as to passing and failing unless these tests have been standardized. An informal test can be diagnostic in the sense that it shows a pupil what he knows and what he does not know. An informal test can tell a pupil his achievement relative to others in the class and he can estimate his chances of passing or failing from this position. Camp[1] suggests that passing shall mean a "mark standing for the teacher's judgment that the pupil has done creditably enough both in amount and quality to progress or to take work in advance, but no better." This definition, which is a very common one in practice, is subjective in two ways. The teacher must be the sole judge as to what is satisfactory and the assigned task may be sufficient or insufficient to meet recognized standards for the course. Judgments about *passing* and *failing* must continue to be troublesome for the teacher with the present testing instruments and methods. Judgments of passing and failing will not be more exact until standardized tests are made available.

[1] Camp, F. S. — "Some 'Marks': An Administrative Problem," *Social Review*, 25:697–713, Dec., 1917.

CHAPTER XXIII

ABILITY GROUPING

Education is famous for its differences in practice and for the acrimony with which the opposing camps discuss the practices of the opposite side. One of these controversies is over ability grouping. Ability grouping, which is a natural result of the application of the findings from measurements to school practice, is opposed on the basis of a variety of arguments. The arguments for and against have been ably expounded and defended or criticized by Briggs,[1] Freeman,[2] and others. The present discussion will first state clearly what is meant by ability grouping and then review the arguments for and against. Following this the basis on which this sectioning should be accomplished will be discussed. Finally, the administration of ability grouping will be briefly touched upon.

Ability grouping is the segregation of pupils studying the same subjects into sections or classes as nearly homogeneous in ability as possible. This grouping is only possible in large schools where there are several sections of the same subject. For instance, if there are 75 pupils in a school who are to study general science, these pupils would naturally be separated into three classes or sections of about 25 pupils each. If these classes are formed by choosing pupils in a random manner, such as by the alphabetical

[1] Briggs, T. H. — "Provisions for Abilities by Means of Homogeneous Grouping," *Third Yearbook of the National Association of Secondary School Principals*, pp. 53–62; 1919.

[2] Freeman, F. N. — "Sorting the Students," *Educational Review*, 68: 169–174, Nov., 1924.

order of their names, three classes will be formed which will be like diagrams A, B, and C in ability. To section on the basis of ability, those with the highest ability would be placed in one section, those with average ability in another and those lowest in ability in another, as expressed by diagrams X, Y, and Z. The exact meaning of ability will be defined later.

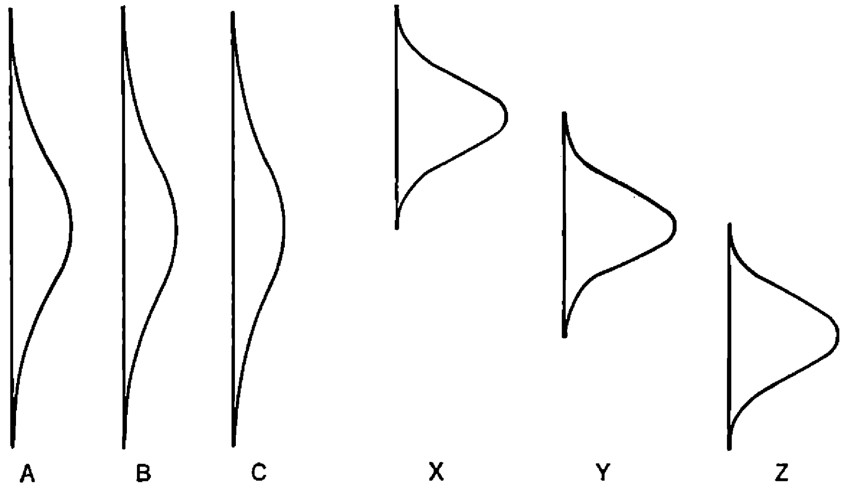


FIGURE 15. — Two methods of grouping pupils

The need for ability grouping has been clearly exhibited by the results of tests of individual differences. Thorndike in his original *Educational Psychology* and in his little book on *Individual Differences* was one of the first to show the great extent and true significance of individual differences. Courtis and Starch [1] were also pioneers in bringing the facts of individual differences strongly before the educational public. The distribution in Table 120 is an illustration of the way children differ in general science. It is typical of distributions found in every high school subject which has been measured. The same sort of distribution is found in intelligence tests.

[1] STARCH, D. — *Educational Psychology*, ch. 3.

TABLE 120

RUCH-POPENOE GENERAL SCIENCE TEST: DISTRIBUTION OF SCORES (JUNE)

SCORES	FREQUENCY
72–75	2
68–71	5
64–67	11
60–63	11
56–59	24
52–55	42
48–51	36
44–47	38
40–43	35
36–39	41
32–35	29
28–31	20
24–27	29
20–23	20
16–19	7
12–15	2
	N = 352

With such wide differences in ability to be found in all groups of children in high school, the plan of ability grouping at once suggested itself to investigators as a means of making provision for these differences.

ADVANTAGES AND DISADVANTAGES OF ABILITY GROUPING

No sooner had the plan to segregate pupils of like ability been suggested than vigorous opposition was expressed. One of the most significant of these arguments is best summed up under the title "educational determinism." This phrase was first used with telling effect by Bagley in a paper on "Educational Determinism; or Democracy and the I Q" read at the meeting of the

Department of Superintendence of the N. E. A. at Chicago in 1922.[1] In this paper Bagley charged that by using tests to effect a form of pupil placement we were thereby insofar controlling pupil educational destiny. If we treat a boy as though he had a high I Q, all is well and good,—he is placed where he can make the most of himself. But if an intelligence test brands a boy as having a low I Q, he is forthwith placed where his educational opportunities are limited by the current practices in the education of this group of pupils. For instance, Freeman [2] suggests that "it might be the policy of a high school to recommend no students for college who have habitually been in a lower section." Segregation on the basis of ability means differential educational treatment and the very act of placing a pupil in one or another section makes for a fatalism or determinism in his education. This argument against ability grouping is really a distrust in both the reliability of tests and the consistency of performance of individuals. This argument can be met by concrete evidence. One line of evidence is the constancy of the I Q. Rugg and Colloton,[3] in summarizing studies of retests of the Stanford-Binet test, find that the average difference in I Q between first and second testings is approximately less than 4.5 points for half of those tested; while Gray and Marsden,[4] in summarizing studies on the Burt-Binet test, report that the middle 50 % of changes in I Q lies between the limits of 5.1 points decrease and 6.0 increase. These results point to a relative steadiness of the I Q.

A second line of evidence comes from the predictive value of tests. These correlations (summarized in Chapters 17, 19, and 20) present an accurate picture of the accuracy of tests for section-

[1] *School and Society*, 15 : 373–384, April 8, 1922. [2] *Op. cit.*

[3] Rugg, H. O., and Colloton, C. — "Constancy of the I Q as Determined by Retests," *Journal of Educational Psychology*, 12 : 315–322, Sept., 1921.

[4] *British Journal of Psychology*, General Section, 15 : 169–173, 1924.

ing. The writer[1] determined the accuracy of certain tests for sectioning in the elementary school from the standpoint of test reliability and found that the "National Intelligence Test," forms *A* and *B*, would misplace 29% and 23%, respectively, in dichotomic sectioning. "These errors in dichotomic sectioning are not all of the same significance. A certain individual, say with an I Q of 110 on a first testing, might, by chance, have a real I Q less than 100. In this case there is a grievous error in sectioning. But another individual might obtain an I Q of 101 on a first testing when his real I Q was 99. Here, again, he would be placed in the wrong section if the dividing line was 100 I Q. Yet this is a very mild error. This latter individual lies so near the border line that it makes little difference in which section he is placed." Woodyard[2] has shown that the constancy of test results is the same whether the test is repeated over a short interval of a day or over a longer interval of several months.

A third line of evidence is that from the results of actual experiments in ability grouping. Reports of measurements which may be considered crucial are lacking. No one has reported the permanency of placement in ability grouping when carried out on a large scale. Harper[3] reports the results of ability grouping in a junior high school in Xenia, Ohio. Intelligence tests were given in June to sixth- and seventh-grade graduates. The groups were reclassified at the end of two months and four months, respectively, on the basis of their ability to do school work (school marks?). The amount of shifting is given in the following table:

[1] Symonds, P. M. — "The Accuracy of Certain Standard Tests for School Sectioning and Marking," *Journal of Educational Psychology*, 15: 423–432, Oct., 1924.

[2] Cf. *The Effect of Time upon Variability*, p. 54.

[3] Harper, M. M. — "Intelligence Tests as a Basis for Homogeneous Grouping," *Elementary School Journal*, 22: 781–782, June, 1922.

TABLE 121

SHIFTING AFTER RE-GROUPING — INDICATING PERMANENCY OF PLACEMENT IN ABILITY GROUPING

	SEVENTH GRADE		EIGHTH GRADE	
	1st Re-grouping	2nd Re-grouping	1st Re-grouping	2nd Re-grouping
Remained in the same group	51%	73%	56%	77%
Moved to a high group . .	24	16	24	12
Moved to a lower group . .	25	11	20	11

In some respects this seems to be a rather large amount of change. Part of the change in the first re-grouping may be due to the basis of the original grouping, which was the I Q's on the intelligence tests given, whereas the first re-grouping was made on the basis of school success, an entirely different sort of measure. In the second re-grouping one must remember that one is checking grouping by one fallible measure against grouping by a second fallible measure. This is not the same as checking grouping by one fallible measure against the true status of these individuals. Notwithstanding the apparently high amount of change found necessary in this school, the amount of change is much less than a purely random change would be.

A second argument against sectioning is that pupils recognize the distinctions made by being placed in different sections and are given a consciousness of superiority or inferiority, which is not wholesome and is often harmful. This is true and cannot be disguised by calling the sections X, Y, and Z instead of A, B, and C. Those who criticize ability grouping on this ground are idealists who would create a situation in school that is not like that found outside of the school and which cannot train pupils to face the situations they will meet after school. As Freeman [1] says,

[1] *Op. cit.*

"Neither the test nor the classification creates distinctions." Ability grouping in school may be used as a means for creating right attitudes toward distinctions that must be faced all through life. In one sense ability grouping lessens rather than increases the consciousness of distinctions. Where a class contains a wide range of ability the dull pupil is made more aware of his deficiencies than when he is among others more like himself in ability. In the latter case he may be given tasks to do with which he can succeed, whereupon he will be less likely to think of himself as a failure than when he is in constant competition with more gifted minds. After all, it is more important for any one to have the consciousness of successful achievement, no matter how humble the task which is undertaken, than to be labeled a failure. The consciousness of inferiority resulting from failure is more serious than that resulting from position. Although what we have been saying has been with reference to the dull, similar statements apply to the bright child, and perhaps with more force. It is more important that a bright child be forced to gain a feeling of superiority by really superior accomplishment than by an almost unearned result of his native ability.

Ability grouping need be used during part of the school day only. Briggs [1] suggests that the pupils be thrown together in the gymnasium, music classes, assembly, etc., wherever the types of abilities used for segregation are not important.

Ability grouping has also been criticised as being undemocratic. In the recent discussion since the war of the significance of democracy with respect to individual differences, it has developed that a democratic state does not want and cannot exist with the assumption of absolute equality. A democratic society requires differences of talent and ability as well as a certain community of ideals and expression. The democratic state in its education must conserve and utilize the native differences of its community.

[1] *Op. cit.*

Freeman [1] says, "The only equality which exists is an equality in the right to opportunity, and opportunity is conditioned by the individual's own nature." The critic of ability grouping is apt to regard it only as a restriction of opportunity for the lower group; he is apt to neglect the fact that for the upper group it means provision of opportunity to progress at their normal rate.

A fourth objection to segregation is based on the belief that the dull need the stimulus of the bright. Interest in school does not come by desire to surpass the bright boy of the class, but by having interesting things to do which are within one's range of ability. If everyday athletes desire competition, they do not race stars. There is more genuine opportunity for competition in classes formed by ability grouping. With homogeneous classes the bright can compete with those who can actually give them a struggle, and the dull have a chance of winning a classroom race or two.

A fifth objection to ability grouping is that gifted pupils will tend to overwork when grouped together. To this objection one may reply that no such malady has been reported where ability grouping has been tried. As a matter of fact, since gifted pupils are much underworked under the traditional organization, there is much leeway before there is a danger of physical or mental strain for the gifted.

A sixth objection is that pupils often learn as much from the recitation of bright pupils as from the explanations of the teacher. If segregated into "C" classes, these slow pupils lose this additional opportunity to learn. Before this objection can be taken very seriously it will have to be demonstrated objectively that pupils learn much from one another's recitations. This point of view is based on an obsolete conception of learning which supposes that learning is a give-and-take activity. As a matter of fact, if slow pupils are stimulated to greater controlled activity when segre-

[1] *Op. cit.*

gated, then they learn even more than in the presence of brighter children who may set an example.

All in all, these objections betray a defective conception of the educative process. The remarks which have been made in reply to these objections have been arguments in favor of ability grouping. The reasons for ability grouping may be summed up under two heads. Ability grouping, in the first place, makes the teaching process easier and hence more effective, and in the second place, it facilitates pupil learning. One has only to note the variability in any distribution of test scores to realize the teacher's task, for this variability represents real differences in learning. At one end are pupils who learn with ease, who know the answer to a question as soon as the question is propounded, and who need to be told how to do a thing only once. These pupils crave original tasks and will willingly follow a topic far beyond the confines of the textbook. The teacher finds it difficult to keep these pupils busy, and if they are live, healthy young people, they will find mischief to do. On the other hand, there are pupils who will occupy five, ten, or fifteen minutes in a recitation and who need much patient help over the difficult places. And then these same pupils will come back the next day and need the same help all over again. I sometimes ask my classes where a teacher should give the most of her time and attention. Some teachers do not realize the problem, for they insist that they can give their attention to all impartially. Some teachers say that they try to fit the needs of the average pupil, others say that they help the slow pupils most because they need it most. Very seldom does a teacher suggest that the recitation should be made to fit the above-average pupil, and yet there is good reason why this should be done. After all, in many of our high school classes the real benefits from the instruction accrue only to the brighter members of the class. This is a hard philosophy for the teacher who is judged by her success with all her pupils, but the situation is one

that the school administrator should be sensitive to. This discussion which draws on the experience of every teacher usually brings home in a decisive way the need of ability grouping from the standpoint of the teacher.

I suppose the teacher would not be considered if the device of ability grouping were not also a help in the learning of the pupils. One is sometimes asked if there is any objective evidence that ability grouping is a success, that is, if it results in accelerated pupil progress. Evidently much was promised to school men by the propounders of ability grouping, for Briggs [1] says that "there has unfortunately sprung up a slogan of 'two terms' work in one' or 'three years' work in two.'" The A sections can and usually do make better than normal progress, but not without endeavor. The A sections will not accelerate their learning unless more is offered them in the way of reward and unless something is gained by their effort. Increased learning does not take place automatically as the result of the organization. B sections usually continue to learn at their past rates. C sections may be expected to progress more slowly. If the slow learners are collected together in a group, they cannot keep up the pace of the average class nor can they be submerged among others. If our schools were run by Squeers to-day, we might expect to have such things hidden, but now if slow-learning pupils are in our schools, we should admit the fact and try to adjust the work to their capacities.

What is the evidence concerning the progress of homogeneous groups? Two studies, neither of which is conclusive, report measures of achievement in classes which have been sectioned on the basis of ability. Cook [2] reports results of homogeneous

[1] *Op. cit.*

[2] Cook, R. R. — "A Study of the Results of Homogeneous Grouping of Abilities in High School Classes," *Twenty-Third Yearbook, National Society for the Study of Education*, part i, pp. 302–312.

grouping in the Topeka High School, Topeka, Kansas. His report is not as valuable as it might have been if he had reported his data on findings in quantitative terms. His conclusions show that the results are not uniform.

Moyer,[1] in an excellently conceived study, compares the achievement of high school pupils who were grouped into classes according to their standing in intelligence tests with pupils of like ability who were taught in mixed classes. The "National Intelligence Tests," the "Hotz Algebra Scales," and "Henmon Latin Tests" were used. Although the number of pupils in each of the groups is nowhere definitely stated, Moyer reports that each of the equal ability sections had an average of 15 pupils; and if this is the total number of cases in a group from which comparisons are made, the results cannot be very conclusive. In the algebra classes segregated superior pupils did no better than unsegregated superior pupils, but the Latin segregated pupils made better scores than unsegregated superior pupils. In the other groups the scores were about the same. Moyer concludes,

> In order to judge the value of segregation we must compare the achievements of pupils who are in segregated classes corresponding to their ability with similar pupils in mixed classes. When we make this comparison, we find that the superior pupils and the medium pupils appear to do somewhat better in segregated than in mixed classes. The inferior pupils, however, do not appear to benefit by segregation so far as their scores on the standard tests are concerned. Whether or not they benefit in other ways the data do not indicate.

Theisen,[2] in a study of ability grouping in the seventh grade, measured and reported the results in objective terms. He found that the brighter groups made better progress than the duller

[1] Moyer, E. L. — "A Study of the Effects of Classification by Intelligence Tests," *Twenty-Third Yearbook of the National Society for the Study of Education*, part i, pp. 313–322.

[2] Theisen, W. W. — "The Relative Progress of VII B Groups Sectioned on the Basis of Ability," *Journal of Educational Research*, 5 : 295–305, April, 1922.

groups. "The average class of highest intelligence was in no case less than three semesters ahead of the average class in the lowest intelligence group." On the other hand, there was considerable difference in the achievement of classes having the same ability. Theisen says, "Grouping by intelligence or mental maturity does not eliminate the necessity of good teaching to bring out the best results. Equal or approximately equal intelligence does not mean equal scholarship. Some low intelligence classes did better than others of high intelligence." However, Theisen is not able to tell us how much better or worse achievement is made by pupils after being placed in homogeneous groups than before being so classified. Careful and large-scale experiments are needed to be able to appraise the results of ability grouping.

The Basis of Grouping Pupils

Once it has been decided that it is advantageous to segregate pupils of like ability, there arises the problem of defining the exact basis on which this sectioning should be made. There are four bases possible for homogeneous sectioning: (1) present status in the subject, (2) present general ability, (3) predicted status in the subject, (4) rate of learning.

By present status is meant the present achievement of a person in a subject. In such subjects as algebra or biology the initial status is zero or so nearly zero as to be insignificant for the purposes of sectioning. But in English, the foreign languages, and in the second term of all classes, achievement is above zero and could be made the basis of sectioning. One would use the scores on an achievement test in order to section on the basis of present status.

By present general ability is meant the score on an intelligence test. Pupils starting a subject for the first time do not have achievement in that subject which may be used as a basis for sectioning, but they may be sectioned if desired on the basis of a general intelligence test.

By predicted status is meant the achievement at the end of the term or year or at the end of the period during which the proposed sectioning plan is to be in force, as measured by a prognosis test.

By rate of learning is meant the rate or speed with which progress will be made in the subject in question. The I Q may be used as a measure of this rate of learning since the I Q measures the rate with which pupils have already made their mental growth or development. It has been demonstrated that there is considerable correlation between the I Q and the rate of learning in any specific function or subject.

Which of these four bases is the proper one to use in ability grouping? In order to answer this question one must know clearly what ability grouping is expected to accomplish. To answer this let us turn back to our description of the nature of growth in the learning of a high school subject. At the beginning achievement is zero — all start at scratch. As time passes some learn faster and some more slowly until at the end of the term or year there is considerable spread in the abilities of the class. This is a familiar phenomenon to every teacher. *The aim of the sectioning is to bring together pupils who will be alike in achievement at the end of a period of learning.* Using this as a guiding principle, "predicted status" would be the basis on which to section. For this purpose prognosis tests are needed in every subject in which it is proposed to section. In the chapter on prognosis tests it was shown that a prognosis test is a test with material selected to predict the particular function in question. Oftentimes a general intelligence test predicts standing in a subject nearly as well as a specially constructed prognosis test. So "present general ability" may also be a basis of sectioning and in lieu of a prognosis test one may use scores on a general intelligence test. It should be understood that a general intelligence test does not fit the situation in a special subject as well as does a special prognosis test. For

instance, with regard to the situation in Detroit, Washburne[1] says,

> "The studies show that excellence in one subject does not prognosticate excellence in others — in 80% of the cases studied, in all ability groups, each child was found at the end of a semester's instruction to be at three or four, out of a possible five, different levels of achievement in the five subjects measured; only 8 per cent of the children maintained a constant position in all subjects."

Notwithstanding this, until prognosis tests have been constructed, intelligence tests may be used for sectioning with satisfactory results.

It may be argued that since final status in the subject should be the basis for sectioning, one should use the I Q rather than the score on an intelligence test. The I Q measures rate of learning and sifts those who will make the most rapid progress from those who will make the least rapid progress. The absurdity of this argument is not apparent until an extreme illustration is taken. Suppose all pupils with I Q's above 100 are to be placed in one section and all with I Q's below 100 in another section. Then one might place a five-year-old boy with an I Q of 110 with another boy fifteen years old with an I Q of 110 in the same section. This extreme situation shows that before the I Q can be used for sectioning there must have been a preliminary sorting on the basis of ability and age. Before one may section using the I Q as a basis, the pupils must represent a group with a certain degree of homogeneity in preparedness for the work in virtue of having passed the preceding grades in school. If they had not already been partially selected on the basis of level of ability by being promoted, one could not use the I Q for sectioning. The order of pupil placement is: first classify on the basis of *level of ability*

[1] WASHBURNE, C. W. — "Adapting the Schools to Individual Differences," Part II, *Twenty-Fourth Yearbook of the National Society for the Study of Education*, p. 165. 1925. Public School Publishing Company, Bloomington, Illinois.

and then section those of like level of ability on the basis of *rate of learning*. Unless pupils have been selected as being of equal initial ability the I Q is of little use in predicting final ability. In using the I Q to predict final ability one is using two entirely different units, one a scale of ability from zero up, the other a ratio with the average at 100. Unless some homogeneity in initial ability has been made, the I Q will not predict final status so well as will the scores themselves on the intelligence test.

This extended discussion of the use of the I Q for sectioning has been made because school practice is so overwhelmingly in favor of using the I Q for sectioning. At every hand one finds school principals turning their intelligence test data into I Q's and using the I Q as a basis for ability grouping. Indeed "section on the basis of I Q" has almost become a slogan. The reason for this is that the procedures for sectioning were first worked out for the elementary school where the use of the I Q for sectioning is quite justified. In the elementary school the first step in classification is grading or bringing together in a grade pupils of like ability, and for this mental age is frequently used as a unit. After pupils of like ability have been brought together, they may be *sectioned* within the grade on the basis of rate of learning, using the I Q. In the high school conditions are different. The first need in pupil sectioning in the high school is to bring together pupils of like *ability*, using *scores* on an intelligence test, achievement test or prognosis test as a basis. If the school is a large one so that there still may be several sections within a single class having the same level of ability, then these might be divided into separate classes on the basis of the I Q.

A, B, and C in figure 16 represent three methods of sectioning 250 pupils in a subject into ten sections of 25 pupils each by means of an intelligence test. In method A one makes ten sections representing ten levels of ability, using scores on an intelligence test as the basis for sectioning. In method B level of ability is

disregarded (except as the level of ability has already been made somewhat homogeneous by virtue of previous promotions) and sectioning is made on the basis of rate of learning, using I Q to determine this. In method C pupils are first divided into three sections representing three levels of ability and these three sections are again subdivided into classes on the basis of I Q. Method C is the one to be preferred of the three methods.

When one is sectioning classes in the middle of the year, or classes such as English or second-year foreign language, several new factors complicate the problem. It was said some pages

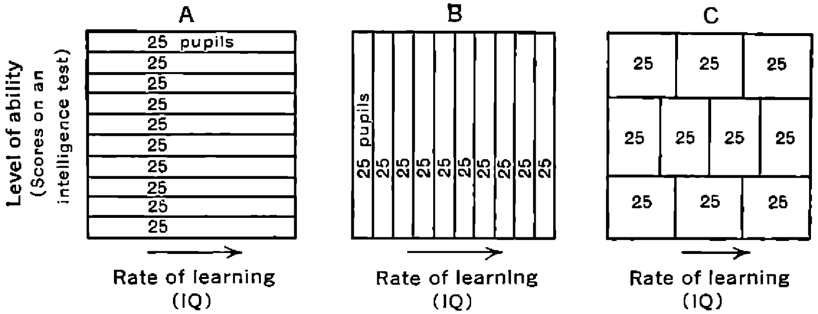


FIGURE 16. — Three methods of grouping 250 pupils into 10 sections

back that the purpose of sectioning was to make the teaching process easier and to facilitate pupil learning. If the purpose is to make the teaching easier, then pupils should be sectioned on the basis of their present achievement, as then they are most homogeneous. If the purpose is to facilitate pupil learning, then they should be grouped on the basis of estimated achievement at the end of the term. To determine this one may use a prognosis test, an intelligence test, or an achievement test. Probably there is no better way of predicting final achievement in a subject than to use present achievement in that subject.

Summary. — The preceding discussion may be briefly summarized as follows: In sectioning classes in high school which are

beginning a subject, use scores on an intelligence test. For those subjects in which adequate prognosis tests have been constructed these prognosis tests are to be preferred to intelligence tests. Do not use the I Q as a basis for sectioning until groups have been made homogeneous as to level of ability. For classes which are not beginning a subject use an achievement test for ability grouping.

Administration of Ability Grouping

There are certain problems of administration in the formation of homogeneous groups that are troublesome. One question relates to the best time for administering the tests. Usually it is best to give intelligence tests during the first or second meeting in the new term. Arrangements should be made for scoring the tests speedily so that the sections may be formed without loss of time. This is generally better than giving the tests at the end of the previous term, because so many changes occur in the student body of a school between sessions. There is no reason, however, why achievement test scores of the previous term should not be used, providing all of the pupils have been given the same test.

Another problem relates to the number of sections. The more sections (and hence smaller sections) a school can afford, the better. There is no magic in having three sections — two, four, or more, sections are just as good. Perhaps it is not best to have more than three sections according to plan A above. If four sections are to be formed, three levels of ability should be found and the middle level of ability might well be divided on the I Q basis. Where three sections are to be formed one might well have them all of the same size. There is a curious superstition abroad that B sections should be larger than A or C sections because of some relationship to the normal curve. Cook[1] says, for instance,

[1] *Op. cit.*

Those persons whose scores were above the middle thirty-eight per cent were placed in the A classes; those whose scores were below the middle thirty-eight per cent were placed in the B sections; those falling within the middle thirty-eight per cent were assigned about equally to the A and B sections. The middle thirty-eight per cent has been selected to represent pupils of medium ability in this division because, ideally, that quantity represents the per cent of cases who should, under a normal distribution of grades in a five-point grading system with limits 2.5 sigma, have received the medium grade.

The normal curve of frequency does not have any arbitrary division points and one is free to draw the lines of demarcation wherever he chooses. If the tests which are given are sufficiently discriminating, one may divide the group so that the numbers in each section are equal or in any other way administratively desirable. There are good reasons for having the A section larger than the C section.

The more often sections are changed, the better, because in that way pupils can be kept in the group that most nearly matches their achievement. It usually happens, however, that three homogeneous sections cannot be kept together. They become fast, average, and slow moving groups. At the end of a term, if the teaching has attempted to fit the abilities of the class, the classes are wide apart. This prevents the interchange of pupils from one section to another after the division is once made. Principals should plan to make three separate tracks, letting each section run at a different rate. If properly managed, articulation points can be found where transfers may be made without loss to any pupil.

In diagram 17 is represented three years' work in language. The normal rate of progress as prescribed by the course of study is represented by section B. Section A does the work of four years in three years. One section starts at the same time as section B, one a half year later, and one a year later. At the end of one half year and one year transfers may be made from section

B to section A_1, and vice versa, without much loss. At the end of one and a half years, two years, two and one half years, and three years, transfers may be made from section B to section A_2, and vice versa, without much loss. Usually when a pupil is transferred to a lower section, he should not be allowed to skip any of the work, as the fact that he is in a section which moves too fast for him is evidence that a review of previous work might be beneficial. When a pupil is moved to a higher section, a small amount of ground may be skipped if necessary, as the evidence which would

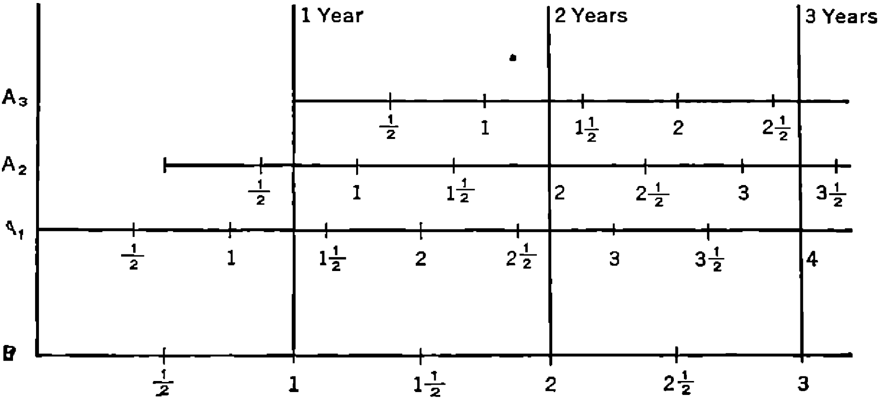


FIGURE 17. — Progress of an average (B) section; and three fast moving sections (A_1, A_2, and A_3) starting at three different times

cause one to place him in a higher section is also evidence that he has exceeded in some way the requirements of the section in which he has been.

Freeman [1] has suggested that only pupils in the middle or upper sections be allowed to take advanced work in any field. This is a wise suggestion, provided plenty of opportunity is given deserving pupils to shift to higher sections if their ability warrants. Any such regulation as this ought not to discriminate to the disadvantage of true ability. But this ruling will seldom

[1] *Op. cit.*

need an exception, as those who are in the lower sections will seldom wish to take advanced work.

Another administrative difficulty in the case of the multiple track plan is that dropping out of pupils often causes the classes to become very small in the upper years so that consolidation becomes necessary. In that case consolidation should unite two sections (such as B and A_2 at the end of the third year) which have reached about the same point in actual achievement.

One of the outcomes of ability sectioning is the fact that sections of different ability require treatment of a radically different sort. An unanswered question to-day in education is whether homogeneous groups should carry the same work, proceeding at different rates, or whether they should proceed at the same rate but have work of different fullness and richness. Terman has been the main exponent in this country of the procedure allowing classes of differing abilities to proceed at their maximum rate. There has been no one champion of the procedure of enriching the course for bright sections and giving only the essentials to slow sections. Caldwell of the Lincoln School of Teachers College has stood for this policy and Holmes of Harvard has written and spoken in favor of it. Some of the large cities which have adopted homogeneous grouping have also advanced different sections at the same rate, trying to provide a more enriched course for the brighter sections. However, this is not a true dilemma, for probably both horns may contain truth. It is not merely a question of covering more topics in general science for the bright sections. Bright sections in such subjects as civics, general science, physics, chemistry, biology, history, and literature may be expected to cover more topics, to do more extensive reading or individual work in each topic, and to take an interest in the more abstract or theoretical aspects of the topic. In the languages the bright sections ought to really make faster progress. The main thing to keep in mind is that sections of different ability require work varied

to fit their ability, otherwise the advantages of grouping are lost.

The matter of marking homogeneous groups has received discussion in the chapter on "Marks and Marking Systems." A difficulty arises, however, in giving credit to the work of different sections for graduation or for college entrance. Freeman suggests that the lowest section be not accredited for college admission. If the bright sections have completed the course in less than standard time, there may be difficulty in acquiring enough "credits" for graduation, or the fifteen points usually required for college entrance. Perhaps this somewhat arbitrary requirement is in the long run helpful in preventing boys and girls from obtaining admission to college when too young. But the day is coming when the colleges will pay more attention to the quality of high school work and will waive strict requirements of amount of work in terms of units of time. The growing recognition of the usefulness of intelligence tests and the value of school marks as criteria of college entrance will help hasten that day.

REFERENCES

Ashbaugh, E. J. — "Homogeneous or Non-Homogeneous Grouping," *Journal of Educational Research*, 9 : 241–245, March, 1924.

Baldwin, B. T. — "The Grouping of Pupils by Abilities in Elementary and High Schools," *Tenth Annual Schoolmen's Week Proceedings*. University of Pennsylvania Bulletin 23, no. 38. Philadelphia, Press of the University of Pennsylvania, June 9, 1923.

Bagley, W. C. — "Educational Determinism; or Democracy and the I Q," *School and Society*, 15 : 373–384, April 8, 1922; also *Educational Administration and Supervision*, 8 : 257–272, May, 1922.

Briggs, T. H. — "Provisions for Abilities by Means of Homogeneous Grouping," *Third Yearbook of the National Association of Secondary-School Principals*. The Association, Cicero, Illinois; pp. 53–62, 1919.

CLERK, F. E. — "Providing for Individual Differences by Grouping by Abilities. Organization and Practical Working of the Plan," *Eighth Annual Schoolmen's Week Proceedings*, UNIVERSITY OF PENNSYLVANIA BULLETIN 2, NO. 37, Philadelphia; pp. 243–249, June 18, 1921.

COOK, R. R. — "A Study of the Results of Homogeneous Grouping of Ability in High-School Classes," *Twenty-Third Yearbook of the National Society for the Study of Education, Part I.* Public School Publishing Company, Bloomington, Illinois; pp. 302–312, 1924.

COURTIS, S. A. — "Ability Grouping in Detroit Schools," in Part II of the *Twenty-Fourth Yearbook of the National Society for the Study of Education.* "Adapting the Schools to Individual Differences." Public School Publishing Company, Bloomington, Illinois; 1925.

COX, P. W. L. — "Providing for Individual Differences by Means of Grouping by Ability," *Ninth Annual Schoolmen's Week Proceedings*, UNIVERSITY OF PENNSYLVANIA BULLETIN 23, NO. 1; pp. 233–244, 1922.

DITMARS, T. — "Intelligence Tests as a Basis for Classification and Grading," *Education*, 44 : 33–39, Sept., 1923.

FEINGOLD, G. A. — "The Sectioning of High-School Classes on the Basis of Intelligence," *Educational Administration and Supervision*, 9 : 399–415, 467–486, Oct., Nov., 1923.

FREEMAN, F. N. — "Sorting the Students," *Educational Review*, 68 : 169–174, Nov., 1924.

GOBLE, W. L. — "Report on Homogeneous Grouping in Illinois High Schools." *Proceedings of the High School Conference*, 1923. UNIVERSITY OF ILLINOIS BULLETIN 21, NO. 25, Feb. 18, 1924. University of Illinois, Urbana; pp. 55–63, 1924.

GRAY, P. L., and MARSDEN, R. E. — "The Constancy of the Intelligence Quotient — Further Results," *British Journal of Psychology*, General Section, 15 : 169–173, Oct., 1924.

HARPER, M. M. — "Intelligence Tests as a Basis for Homogeneous Grouping," *Elementary School Journal*, 22 : 781–782, June, 1922.

HUGHES, W. H. — "Provision for Individual Differences in High School Organization and Administration," *Journal of Educational Research*, 5 : 62–71, Jan., 1922.

JOHNSON, F. W. — "Classification of Pupils according to Ability," *Contributions to Education*, Vol. I, J. C. Bell, ed.; pp. 214–220. World Book Company, Yonkers-on-Hudson, New York, 1924.

MOYER, E. L. — "A Study of the Effects of Classification by Intelligence Tests," *Twenty-Third Yearbook of the National Society for the Study of Education*, Part I. Public School Publishing Company, Bloomington, Illinois; pp. 313–322, 1924.

NICHOLS, M. L. — "The Teaching of Science to Classes Divided According to Ability," *Ninth Annual Schoolmen's Week Proceedings*, UNIVERSITY OF PENNSYLVANIA BULLETIN 23, NO. 1; pp. 324–326, 1922.

ODELL, C. W. — "An Annotated Bibliography Dealing with the Classification and Instruction of Pupils to Provide for Individual Differences," *Bureau of Educational Research Bulletin*, No. 16. College of Education, University of Illinois. University of Illinois, Urbana; Nov. 19, 1923.

OMANS, A. C. — "Provision for Ability Grouping in Junior and Senior High School," *American School Board Journal*, 65 : 55–58, Oct., 1922.

PECHSTEIN, L. A. — "Homogeneous Grouping of High School Pupils by Intelligence Tests," *Proceedings, Third Annual Ohio State Educational Conference*. THE OHIO STATE UNIVERSITY BULLETIN, 28, Aug. 31, 1923. Ohio State University, Columbus; pp. 352 f., 1923.

ROWLAND, S. V. — "Changes in Educational Procedure Growing Out of the Grouping of Pupils by Ability," *Tenth Annual Schoolmen's Week Proceedings*. Press of the University of Pennsylvania, Philadelphia; pp. 64–69, 1923.

RUEDIGER, W. C. — "Homogeneous Grouping of Pupils," *Eleventh Annual Schoolmen's Week Proceedings*. UNIVERSITY OF PENNSYLVANIA BULLETIN 24, NO. 38, June 7, 1924. Press of the University of Pennsylvania, Philadelphia; pp. 78–84, 1924.

RUGG, H. O., and COLLOTON, C. — "Constancy of the I Q as Determined by Retests." *Ref.* Ch. 15.

SANDBERG, D. — "Educational Practices Growing Out of the Homogeneous Grouping of Pupils," *Eleventh Annual Schoolmen's Week Proceedings*. UNIVERSITY OF PENNSYLVANIA BULLETIN 24, VOL. 38, June 7, 1924. Press of the University of Pennsylvania, Philadelphia; pp. 113–119, 1924.

SEASHORE, C. E. — "Sectioning Classes on the Basis of Ability," *School and Society*, 15 : 353–358, April 15, 1922.

STETSON, P. C. — "Homogeneous Grouping in the First Year of a Five-Year High School," *School Review*, 29 : 351–365, May, 1921.

SUTHERLAND, A. H. — "Ability Grouping in Los Angeles" in Part II of the *Twenty-Fourth Yearbook of the National Society for the Study of Educa-*

tion. "Adapting the Schools to Individual Differences." Public School Publishing Company, Bloomington, Illinois, 1925.

SYMONDS, P. M. — "The Accuracy of Certain Standard Tests for School Sectioning and Marking." *Ref.* Ch. 15.

THEISEN, W. W. — "The Relative Progress of VII–B Groups Sectioned on the Basis of Ability," *Journal of Educational Research,* 5 : 295–305, April, 1922.

TRABUE, M. R. — "The Use of Intelligence Tests in Junior High Schools," *Twenty-First Yearbook of the National Society for the Study of Education.* Intelligence Tests and Their Use, Part II. The Administrative Use of Intelligence Tests, ch. vi. Public School Publishing Company, Bloomington, Ill., 1922.

VOEGELEIN, L. B. — "An Annotated Bibliography on Adapting Schools to Individual Differences," Part II. Classification, Ability Grouping and Promotion Plans, *The Twenty-Fourth Yearbook of the National Society for the Study of Education.* Part II. "Adapting the Schools to Individual Differences." Public School Publishing Company, Bloomington, Ill., 1925.

WASHBURNE, C. W., COURTIS, S. A., SULLIVAN, E. J., WARD, M. A., CARTER, G. E., HOLMES, H. M., ANDERSON, C., HORN, E. "Data on Ability-Grouping from Detroit, Los Angeles, Winnetka, San Francisco, State Teachers College, Iowa, and a Survey of these Data." In Part II of the *Twenty-Fourth Yearbook of the National Society for the Study of Education.* "Adapting the Schools to Individual Differences." Public School Publishing Company, Bloomington, Ill., 1925.

WERREMEYER, D. W. — "The Benefits of Grouping Students in Mathematics on the Basis of Ability," *Proceedings of the Second Annual Educational Conference.* THE OHIO STATE UNIVERSITY BULLETIN 27: 16, March 24, 1923. Ohio State University, Columbus; pp. 300–303, 1922.

WISE, C. R. — "Ability Grouping from the Administration's Viewpoint," *Proceedings of the Fourth Annual Ohio State Educational Conference.* THE OHIO STATE UNIVERSITY BULLETIN 29, NO. 2, Aug. 31, 1924. Ohio State University, Columbus; pp. 284–287, 1924.

WOODY, C. — "Measurement of the Effectiveness of Differentiation of High School Pupils on the Basis of the Army Intelligence Tests," *Journal of Educational Research,* 7 : 397–409, May, 1923.

WOODYARD, E. — *The Effect of Time upon Variability,* TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 216, New York, 1926.

CHAPTER XXIV

MARKS AND MARKING SYSTEMS

In Chapter I of this book a list of fifteen uses of measurements in high schools is given. This list might almost be repeated here without change as indicating the function of marks in school, for almost all measurements in school crystallize themselves in terms of school marks. These marks are the school's record of pupil achievement. School marks find the following specific uses:

1. To inform pupils and parents of pupils' achievement.
2. As incentives to study.
3. To promote competition.
4. To determine promotion.
5. To determine graduation.
6. To predict a pupil's future success.
7. To enable college authorities to pass on the qualifications of entrance candidates.
8. To determine credits, honors, etc.
9. To determine participation in extra-curricular activities.

The first three of these might well be called the pedagogical function of marks. These functions are only too often ignored by teachers who are impatient of the recurring demands made upon them for marks. Some teachers would like to be relieved of the necessity of this quarterly task of filling in the report cards of their pupils. They fail to realize that these marks fill an important function and that without them a substitute (which would prob-

ably be even more subjective than even the present system) would have to be adopted. Pupils require some sort of statement of their achievement if learning is to take place at all. The formal term or quarterly marks become hurdles or milestones which every one finds so necessary in his career as compared with the day-to-day successes or failures. To be sure, the current system of letters or percentages does not meet this need as it merely indicates relative achievement. Nevertheless, one has only to be present in a school on the day on which marks are distributed to feel the intense significance they have for the student body.

An unsolved problem is the extent to which pupils work for marks. It is certain that the day-to-day activities are more immediate, and in many cases most influential, in motivating effect. A pupil works to get the day's lesson and the success in class is probably a foremost incentive to study. But it cannot be denied that longer range objectives are also effective and that many pupils work for marks. Such an extrinsic motive as working for marks is decried by many teachers who wish for an intrinsic interest in the subject matter. These same teachers, however, will come to summer school and work for marks as enthusiastically as any of their own charges. Unfortunately the intrinsic interest is partly a result of learning and cannot be relied on for motivation. One should not be ashamed of this dependence on marks for motivation — here, if ever, the end justifies the means.

The most obvious uses of the school mark are the administrative uses. School marks are used first of all to determine promotion. In this connection the need for accuracy is apparent. A pupil's whole educational career is markedly influenced by promotion or failure of promotion. The inconsistency of the current point of view is well illustrated when a teacher insists that he cannot afford time from his instruction in order to give a reliable test, when the result of this test would determine the placement

of several of the pupils of the class, making a difference of several months in their ultimate educational advance. The importance of accurate placement was not demonstrated until it was shown that of all the factors influencing the A R, grade placement was the most important. Teachers will probably only take a keener interest in assigning more accurate marks when they realize the importance of accurate pupil placement. Longer and more adequate tests will pay liberal educational dividends.

In other chapters it has been shown that school marks are one of the most important predictions of future school success. In order to be able to give pupils adequate guidance, marks should be as accurate as possible.

More use should be made of school marks than is being made at the present time in most schools. School record cards should be so frequently consulted as to become dog-eared. Every room teacher, where the room teacher acts in the capacity of adviser, should have and use the complete record of every pupil in her room. A teacher should refer to the marks made by her pupils in other classes. Consistently high or low marks made by a pupil are an excellent symptom of ability. The marks also show special interest and abilities which oftentimes help explain certain instances of disparity between intelligence and achievement. The supervisor of a study hall should have at hand the record of each pupil studying in the room in order to know whom to help and what sort of help to give.

On the next page is a scheme for giving absolute marks. This scheme cannot claim perfection. It is based upon only 8 standardized tests and these tests have limitations. Such a scheme cannot take care of curriculum variations or of changes in the course of study. Perhaps these limitations can be overcome in some way. At least this marking scheme is based upon absolute standards, and its use, if perfected, would tend to improve marks for the uses to which they are put.

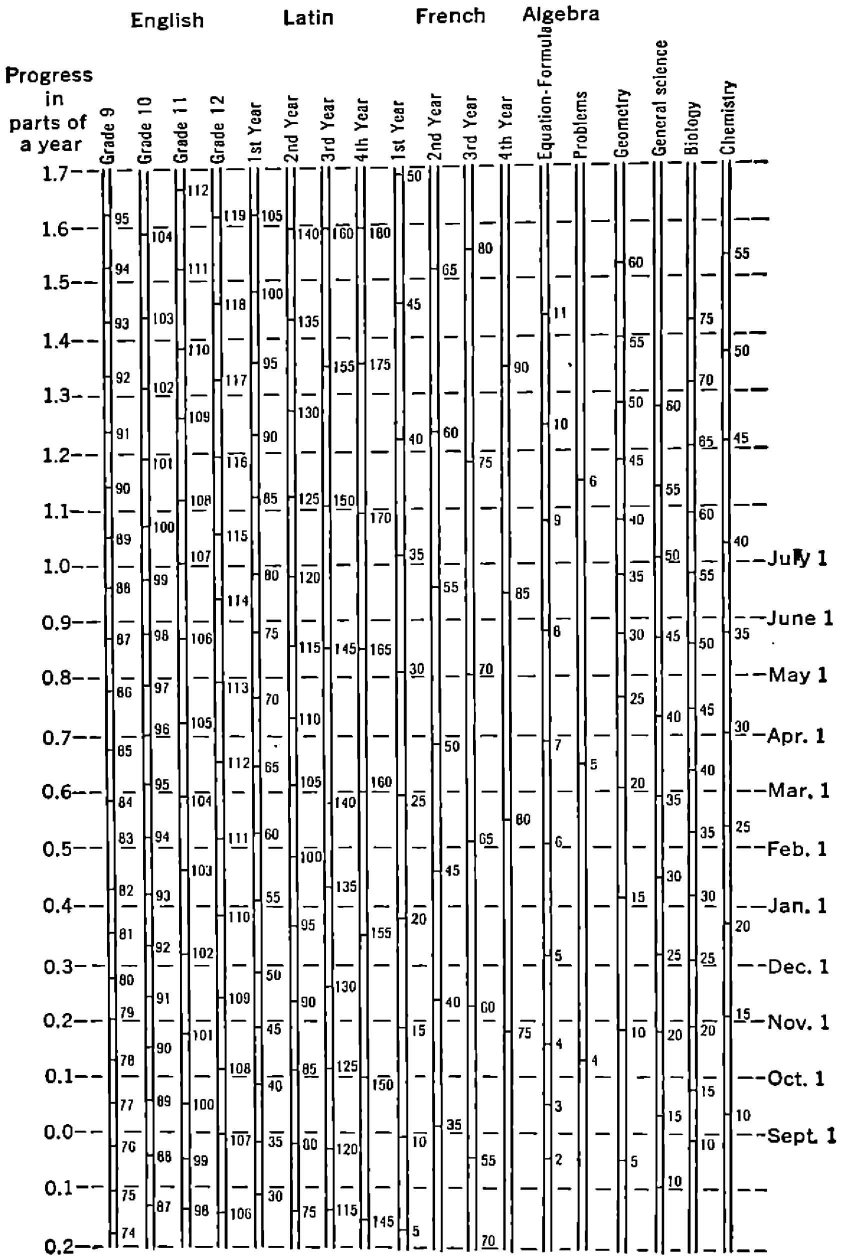


FIGURE 18. — Showing progress on eight standardized tests throughout a school year: a suggested form for recording absolute marks

The tests used in figure 18 are:	*Publisher*
Haggerty Reading Sigma 3	World Book Company
White Latin	World Book Company
Standards are given separately for grades 9, 10, 11, 12. These are meant to correspond with 1st, 2nd, 3rd, and 4th years of Latin respectively; or Beginning Latin, Cæsar, Cicero, Virgil, respectively.	
American Council French Test — Part I	World Book Company
The standard represents the total on the vocabulary and grammar tests combined. Standards are given separately for 1st, 2nd, 3rd, and 4th year French.	
Hotz Algebra Scales — Series A	Teachers College
Equation — Formula test	Bureau of Publications
Problem test	
(Separate standards are given for the scores on these two tests.)	
Schorling-Sanford Geometry Test	Teachers College Bureau of Publications
Ruch-Popenoe General Science Test	World Book Company
Ruch-Cossman Biology Test	World Book Company
Powers Chemistry Test	World Book Company
Van Wagenen American History Test S-3	Teachers College Bureau of Publications

Since the plan above described has yet to be perfected, the present marking systems must continue to be used. What are they? Odell[1] surveyed the marking systems of the State of Illinois and found among 281 schools almost 100 varieties. These marking systems are displayed in the following table.

About three fifths of the schools using letters or figures define the letters or figures in terms of percentages. It is seen from this

[1] Odell, C. W. — "High School Marking Systems," *School Review*, 33: 346–354, May, 1925.

TABLE 122

MARKING SYSTEMS IN USE IN 281 ILLINOIS HIGH SCHOOLS

KIND OF MARKING SYSTEM	NO. OF SCHOOLS	PER CENT
Percentile (marks given in actual percentages)		
Pass at 70:		
No conditions	40	14.2
Conditions at 60, 65, 67, 68, 69, or indefinite	16	5.7
Pass at 75:		
No conditions	81	28.8
Condition at 70	38	13.5
Conditions at 60, 65, 72, 73, 74, or indefinite	30	10.7
Pass at 80:		
No conditions	1	.4
Total	206	73.3
Letters or figures (*A*, *B*, *C*, etc.; *S*, *E*, *G*, *F*, *P*, etc. 1, 2, 3, etc.):		
3 passing, 1 failing, marks:		
Without + and −	8	2.8
With + and −	5	1.8
3 passing, 2 failing, marks, without + and −	1	.4
3 passing, 1 conditioned, 1 failing, marks with + and −	14	5.0
4 passing, 1 failing, marks:		
Without + and −	22	7.8
With + and −	3	1.1
4 passing, 1 conditioned, 1 failing, marks:		
Without + and −	4	1.4
With + and −	6	2.1
4 passing, 2 failing, marks, with + and −	2	.7
5 passing, 1 failing, marks, without + and −	3	1.1
5 passing, 1 conditioned, 1 failing, marks:		
Without + and −	5	1.8
With + and −	1	.4
7 passing, 1 failing, marks, without + and −	1	.4
Total	75	26.7
Grand total	281	100.1

that the percentage system of marking dominates. The percentile values of letters are usually expressed as ranges; for example, *A* may mean 90 to 100; *B*, 80 to 89; *C*, 70 to 79; etc. Occasionally, however, letters are given definite percentile meanings. Thus, *A* may equal 95; *B*, 85; and *C*, 75.

The rather frequent use of *conditions* seems to the writer to be a tacit admission of the inaccuracy of marking. One meets teachers who would never give a mark of 59% or 69% or 79%, claiming that they are not sure enough of themselves in their marking. If ever their calculations in arriving at marks should lead to a 59 or 69 or 79, they would call it 60 or 70 or 80, believing that the pupil is entitled to the benefit of the doubt which exists in their own minds. These same persons insist that no one else could mark any more accurately. It is because of a general belief of this sort that permits a condition at 67 or 68 or 69 which can be made up subsequently by the pupil. This is an open admission on the part of school authorities of the inaccuracy of school marks.

On What Should Marks Be Based?

Part of the confusion regarding marks is that teachers determine marks on all sorts of different bases. Some years ago Camp[1] held faculty meetings in his school in Stamford, Connecticut, to discuss marks. The topic of the first meeting was "What We Mark." The discussions brought out the fact that teachers mark, among other qualities, improvement, ability, serious purpose, moral qualities, interest in work, accomplishments. The domestic science department marked "conscientious work, improvement, and ability." The commercial department marked "accuracy, neatness, and promptness," which by the way were

[1] Camp, F. S. — "Some 'Marks': an Administrative Problem," *School Review*, 25: 697–713, Dec., 1917.

three slogan words very properly adorning the walls of the department. They also considered "honesty, courtesy, and zeal." The history department marked "endeavor and accomplishment." One teacher in the mathematics department offered "acquisition of knowledge." Camp adds that this appalling category contained all that is good and moral.

Let it be stated here with emphasis that school marks should be thoroughbred and not hybrid. *School marks should represent only one thing* — ***achievement.*** School marks must not be a composite of several different qualities, otherwise they mean nothing. Above all, school marks should not be based merely on conscientious work. Quite unconsciously teachers use marks not only as a measure of achievement but as a disciplinary weapon. The good, quiet, industrious pupil is rewarded by giving him high marks, whereas the troublesome, lazy, noisy pupil is penalized by giving him low marks, notwithstanding his achievement. Teachers only too often let their emotions and sympathy come between their calm judgment and the marks they assign. "C—— has worked so hard this term that I just cannot give him a D. He may do better next term," or "I don't see how T—— did so well on his examination. He causes me more trouble in class than any half dozen others put together. He is noisy and inattentive and half the time unprepared. I can't help it if he is bright. It would only be a bad example to the rest in the class to give him an A." So these two boys receive marks not on the basis of what they have done but on the basis of what the teacher hopes they will do. It is difficult oftentimes to mark strictly on the basis of achievement. Perhaps the pupil has been ill during the term and the evidence shows that he has fallen behind the others in the class. His marks should testify to this fact, but the tendency is to excuse the pupil because he was sick and raise his mark accordingly.

I was once in a group of educators where it was seriously argued that pupils in any class should be marked on the basis of an

intelligence test. The argument was that since Franzen[1] had shown that with maximum motivation there would be perfect positive correlation between achievement and ability, it was quite proper to anticipate the event and use the intelligence test as a basis for marking. The fallacy in the argument is that in no actual situation do we have maximum motivation. Marks should be assigned to pupils on the basis of achievement quite independent of intelligence test scores.

A corollary of all this is that one should use objective evidence of achievement. In other words, school marks should be based solely on the results of *tests*, so far as possible. This does not necessarily mean that all should depend upon a final examination, for quizzes may be given frequently throughout the term, but that all the evidence upon which marks are based should be objective in character.

I have seen teachers try to mark conscientiously every pupil recitation. These records of individual recitations must be perforce highly unreliable. Any one who has attempted to rate a recitation is aware of this fact. Usually the teacher has some answer which he wants and which will receive credit while another answer which may be quite as good will not receive credit. Then again such things as timidness, stuttering, awkwardness, clumsiness, poor command of English, and general appearance confuse the impression of the actual recitation.

Some principals who distrust the results of a single examination try to define the mark by requiring that it be based one third on the final examination and two thirds on the daily work. Daily work is usually interpreted as the quality of recitation or a teacher's general impression of the pupil. These two factors, however, should have no place whatever in the assigning of marks. Marks should be based entirely on objective evidence, such as tests. Standardized tests or informal tests, preferably

[1] *Cf. The Accomplishment Ratio*, pp. 55–59.

of an objective type, may be used or, if the tests are not objective, effort should be made to mark them in a consistent manner. Tests need not measure merely the information side of the course. Every teacher should formulate the objectives of the course he teaches and then try, so far as possible, to test these objectives impartially and objectively.

One way in which school marks can be kept clean to record achievement alone is to provide marks for the recording of other qualities. Later in this chapter an objective measure of studiousness or effort will be described. Many schools are now having teachers record their estimates of other character traits on the report card. Although these judgments of character traits possess much unreliability, they at least clarify the issue and make it easier for the marks which record achievement to perform their task without confusion.

Variations in School Standards of Marking

To agree on what the mark represents and to use reliable measures is the first step in making marks accurate. The next problem for the school principal to turn his attention to is the problem of standards of marking among the teachers of the school. The wide discrepancy of standards between teachers in a school may pass unnoticed for many years.

Johnson [1] was the first to note and study extensively this phenomenon. His study is described in Chapter II of this book and the reader may turn to pages 10–18 to find examples of the variations in standards of marking. Kelly,[2] in discussing Johnson's data, does not believe that Johnson's figures are convincing. He believes that even such wide variations as reported by Johnson could be explained by factors other than varying

[1] Johnson, F. W. — "A Study of High School Grades," *School Review*, 19: 13–24, Jan., 1910.

[2] *Cf. Teacher's Marks, Their Variability and Standardization*, 1914.

standards held by teachers. Kelly says, "It is conceivable that a set of conditions might prevail in which the above variations would be justified, at least in part. Johnson offers them, however, as examples of variation which have no justification. They are simply due to different standards held by different teachers."

Not until standardized tests were available was it possible to untangle the factors involved. The following figures were obtained from eight sections in senior French in a large city high school. The second column contains the average scores made by each section on a standardized test in French. The fourth column gives the average mark assigned by the teacher of each section. ($A = 1, B = 2, C = 3, D = 4, E = 5, F = 6$.)

TABLE 123

AVERAGE SCORES ON A STANDARDIZED FRENCH TEST AND AVERAGE MARK GIVEN BY TEACHERS IN EIGHT SECTIONS OF SENIOR FRENCH IN A CITY HIGH SCHOOL

SECTION	AVERAGE SCORE ON TEST	RANK	AVERAGE MARK	RANK
A	80.64	1	4.011	1
B	81.25	2	3.844	2
C	84.33	3	3.812	3
D	89.00	4	3.447	4
E	90.15	5	2.961	6
F	92.25	6	3.145	5
G	94.67	7	2.488	8
H	108.00	8	2.797	7

Although there is good agreement here between the scores on the test and the marks, the agreement is not perfect. One should notice that the variability of the mean of marks (from what corresponds to $C+$ to D), is much greater comparatively than the variability of the achievement test scores, which, with the excep-

tion of Section *H*, are very nearly alike. Weld[1] asserts that "we may expect any one large group of unselected pupils (as those handled by one teacher in a year) to have about the same actual scholarship, on the average, as any other similar large group."

Every school should annually study the distribution of marks in the school and bring the distribution of marks to the attention of the teachers. Such a study might well take the following form.

TABLE 124

SUGGESTED FORM FOR THE STUDY OF A SCHOOL'S DISTRIBUTION OF MARKS
Distribution of Marks (Per Cent)

TEACHER	*F*	*D*	*C*	*B*	*A*
A	2%	11%	25%	40%	22%
B	22	26	21	28	3
C	7	14	42	32	5
D	etc.				
E					
etc.					
Total School	7	16	38	33	6

In order to secure a single figure for each teacher, each mark may be translated into a number, *i.e.*, $A = 5$, $B = 4$, $C = 3$, $D = 2$, and $F = 1$. Then the average mark given by each of the above teachers is

TEACHER	AVERAGE MARK	INTERPRETATION
A	3.69	too high
B	2.64	too low
C	3.14	about right
Total School	3.15	

[1] WELD, L. D. — "A Standard of Interpretation of Numerical Grades," *School Review*, 25: 412–421, June, 1917.

This interpretation by using only the average disregards variations in the variability of marking, but this may be safely disregarded for the present purpose.

Several methods of overcoming the variability in standards of marking have been proposed. Since it is found that teachers acquire a tendency to mark too high or too low, correction formulæ or correction tables may be employed to bring all marks on an equable basis. A famous plan known as the Missouri System of Grading was developed by Max Meyer and put into operation in the University of Missouri. Meyer's scheme was to have each instructor send to the office, not a mark for each student, but a list of the students in rank order of their achievement. These rank orders were turned into per cents and these per cents were used as the basis for marking. If a student stood 20th in a class of 47, his rank order per hundred was $\frac{20}{47} = .43$. This scheme insures uniformity of standards between classes although it robs the instructor of the right to assign his own mark.

The plan which is most widely accepted to insure uniformity of standards is the use of the normal probability curve. Cattell [1] was the first to note that if a large number of teachers' marks were tabulated, the distribution approximated very closely to the normal probability curve. This phenomenon has been studied again and again by such investigators as Meyer,[2] Dearborn,[3] Starch,[4] Finkelstein,[5] Pickell,[6] Rugg,[7] and others and always with the same result. The use of standardized tests in various school

[1] CATTELL, J. McK. — "Examinations, Grades and Credits," *Popular Science Monthly*, 66 : 367–378, 1905.

[2] MEYER, M. — "The Grading of Students," *Science*, 28 : 243-252, Aug. 21, 1908.

[3] *Cf. School and University Grades.*

[4] *Cf. Educational Psychology*, ch. 22.

[5] FINKELSTEIN, I. E. — *The Marking System in Theory and Practice*, 1913.

[6] PICKELL, F. G. — "Credit for Quality in the Richmond (Ind.) High School," *Educational Administration and Supervision*, 3 : 537–538, Nov., 1917.

[7] RUGG, H. O. — "Teachers Marks and Marking Systems," *Educational Administration and Supervision*, 1 : 117–142, Feb., 1915.

subjects has given added substantiation that with a large number of cases the form of distribution approximates the normal curve. Thorndike and Bregman [1] have shown that this is true of ability in general in a grade or age group. This fact is also approximately true in single classes. Accordingly it has been suggested that one way to insure uniform standards in marking throughout a school is to require that the marks of the school and of each class approximate the normal distribution or some standard distribution agreed upon by school authorities. This plan has been subjected to severe criticism by those who wish to have no tampering with their prerogative to assign marks. Such persons are not aware that the marks which they assign are closely bound up with the marks given by other instructors so as to influence the discipline of the school and the selection of courses by the students. This plan has also been abused by trying to enforce it too rigorously. The normal distribution is only a guide — it is not expected that it will be applied to the last decimal place.

In Table 125 is shown percentage distributions for different sized classes and for scales having different number of points.

The first group of percentage distributions is the one to be employed by the classroom teacher. Since the modal system of marks is one having 5 letters, *A*, *B*, *C*, *D*, and *F*, the distribution of *A*, 7%, *B*, 24%, *C*, 38%, *D*, 24%, and *F*, 7% is the one which will be most commonly used. With reliable measures much finer discrimination can be made than that expressed by a five-point scale. There is no reason why a seven-point scale should not be used, such as *A*, *B*, *C*, *D*, *E*, *F*, and *G*, or even one having still finer divisions. To continue to use a five-point scale is to admit of blurred vision and inaccurate marking.

[1] Thorndike, E. L., and Bregman, E. O. — "On the Form of Distribution of Intellect in the Ninth Grade," *Journal of Educational Research*, 10 : 271–278, Nov., 1924.

TABLE 125

PERCENTAGE DISTRIBUTIONS FOR CLASSES OF DIFFERENT SIZES AND FOR SCALES HAVING DIFFERENT NUMBER OF POINTS

CLASS OF 40 (RANGE 5 S.D.)				
3 Groups	4 Groups	5 Groups	6 Groups	7 Groups
20%	11%	7%	5%	4%
60	39	24	15	10
20	39	38	30	22
	11	24	30	28
		7	15	22
			5	10
				4
CLASS OF 185 (RANGE 6 S. D.)				
3 Groups	4 Groups	5 Groups	6 Groups	7 Groups
16%	7%	4%	2%	2%
68	43	24	14	8
16	43	44	34	23
	7	24	34	34
		4	14	23
			2	8
				2

The second group of percentages is the one to be employed by the head of a department in making a composite distribution for several sections of a class in the department. Since it implies a distribution with a range of 6 S. D., the percentages at the extremes of the scale are smaller than for the distributions implying a range of 5 S. D.

The percentages derived from the normal probability curve are only an approximation. In any one class the achievement of the group might exhibit marked deviations from the normal distribution. It is probably safer always to use the distributions

as obtained from a given set of test scores than to force a group to fit the normal distribution.

How to Turn Test Scores into Letter Marks

Although the point of view of this chapter is that the only defensible marking system is an absolute system, such a plan is not possible at present and the best must be made of the letter system of marks. This section describes a method for turning test scores into letter marks so that they harmonize with the other recommendations made in this chapter. The data for the following example were kindly sent me by Mr. Carl Burg of the Stewartstown High

TABLE 126

Scores and Suggested Marks for a Class of Thirty-Six Pupils in General Science

Name of Pupil	Score on Test	Mark	Name of Pupil	Score on Test	Mark
J. C.	43	B	J. H.	27	D
R. D.	24	D	E. J.	26	D
H. E.	27	D	M. K.	43	B
J. G.	56	A	M. L.	31	C
W. H.	23	D	C. M.	32	C
J. J.	38	B	K. M.	28	C
D. M.	40	B	E. M.	20	D
G. M.	28	C	D. O.	38	B
R. M.	45	A	M. O.	42	B
W. T.	42	B	M. R.	27	D
E. A.	21	D	S. R.	25	D
H. A.	40	B	E. S.	25	D
M. B.	33	C	G. T.	27	D
R. D.	45	A	R. T.	30	C
E. F.	22	D	C. W.	26	D
E. G.	22	D	M. W.	25	D
A. H.	30	C	B. W.	38	B
M. H.	31	C	L. W.	37	B

School, Pennsylvania. The following scores are results from giving the "Ruch-Popenoe General Science" test in June.

Step 1: Find the mean and standard deviation of the distribution of scores according to the method described in Appendix A. In this particular example

$$M = 32.1 \quad \text{and} \quad \sigma = 8.5.$$

Step 2: Divide the base of the distribution into intervals of 1 σ as follows:

A	above $+ 1.5\ \sigma$
B	from $+ 0.5\ \sigma$ to $+ 1.5\ \sigma$
C	from $- 0.5\ \sigma$ to $+ 0.5\ \sigma$
D	from $- 1.5\ \sigma$ to $- 0.5\ \sigma$
E	below $- 1.5\ \sigma$

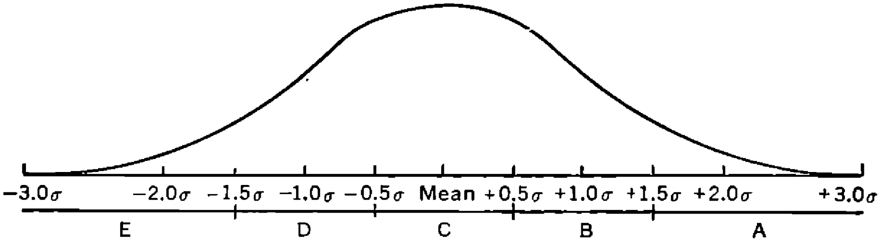


FIGURE 19. — The normal distribution of marks

Step 3: Turn these σ units into score units. $- 1.50\ \sigma$ is obtained as follows: $1.50\ \sigma = 1.5 \times 8.5 = 12.75$. This value is then subtracted from the mean to give $- 1.50\ \sigma$. $32.1 - 12.75 = 19.35$. The limits for the marks are thus transferred into score units.

A	above 44.85
B	36.35–44.85
C	27.85–36.35
D	19.35–27.85
E	below 19.35

or more simply

A	45 or above
B	37–44 inclusive
C	28–36
D	20–27
E	19 or below

Step 4: Assign marks to individuals. J. C. with a score of 43 would be given a mark of *B* and so on with the others. The marks that are assigned individuals may be found in the original table of scores on page 513.

The distribution of marks is as follows:

	No.	%
A	3	8
B	10	28
C	8	22
D	15	42
E	0	0
Total	36	100%

One will note that the above distribution is far from normal. There are no *F*'s although the percentage of *D*'s is quite abnormally large. This distribution, as found, probably represents the true distribution of the class better than an application of the normal distribution would do. Some such procedure as the above is advised wherever standardized or informal new type tests are used. The writer in his classes is in the habit of giving several informal tests during a term and a longer test of the new type at the end of the course, then aggregating the scores without attempting to weight the various tests used. This weighting may be disregarded because the number of total elements is so large that no one of them has much influence on the score. From the distribution of the composite of all tests given, the above procedure is followed in arriving at marks. This method may

be used instead of attempting to follow the normal distribution, and probably with better results.

In some schools there is a general tendency to skew the distribution of marks away from the higher marks, *i.e.*, to give a greater percentage of *A*'s and *B*'s than *D*'s and *F*'s. If this is the case, the procedure in step 2 should be modified. If, for instance, it was found that the distribution of marks for the whole school was as given in the following table and it was not desirable to change these percentages because of the peculiar significance which each mark had acquired in the school, one could use a table of the probability integral to determine the σ limits for the different marks. The precise procedure for this will not be described here.

A	12%	above $+ 1.14\sigma$
B	22	from $+ .40\sigma$ to $+ 1.14\sigma$
C	42	from $- .69\sigma$ to $+ .40\sigma$
D	20	from $- 1.68\sigma$ to $- .69\sigma$
E	4	below $- 1.68\sigma$
	100%	

Equating Marks

One common criticism of the use of the normal curve of frequency is that classes are often far from normal — some may contain pupils who are better than average and others pupils who are poorer than average. This is in some degree true, but it was the abuse of this truth that led to the situation that requires the use of the normal curve of frequency. Without the use of a common test no teacher knows whether or not he has a select group, except by comparing his pupils with those of previous years. Optimistic, enthusiastic teachers with a firm belief in the untouched potentialities of every boy and girl in their classes always believe they have exceptional groups and should be permitted to give high marks; while the pessimistic, overworked teacher who takes his work too seriously and does not have faith in his pupils will wish to mark

too low. Domestic arts and physical education departments are usually inclined to believe that nearly all of their pupils should qualify for *A*'s and *B*'s, whereas the language and mathematics teachers wish to fail a large percentage. Now the truth is that without definite knowledge the safest assumption is that classes are of nearly the same ability. One might expect that classes in intermediate algebra who had survived beginning algebra or classes in physics and chemistry would be select classes, and they are to some extent, but not nearly so much as most teachers would expect. The use of the normal curve probably does little injustice as compared with giving teachers free rein and letting them assign marks without check.

To allow for differences in selection of classes in using the normal probability curve, the following method is suggested which makes use of intelligence tests. Intelligence tests of course are not to be used as a measure of achievement, but they may well be used to determine the general level of any group of pupils. In making this use of intelligence tests, little reference is made to the fact that they are or are not "intelligence" tests. They are used because they represent a measure of ability applied to the whole group. An algebra or general science test would be equally satisfactory if (1) algebra or general science had been studied by all the pupils and (2) all the pupils had taken the same test. Such conditions are fulfilled by an intelligence test. An intelligence test contains material which is fair to all pupils and all pupils may be given the same test. Hence the level of ability of any group of pupils may be found with respect to the total group with a tolerable degree of accuracy.

Following is a method worked out in detail by which this equating scheme can be easily and economically applied. First, determine the distribution of the intelligence scores of the total student body. In high school this should properly be restricted to one class, as freshman or senior. For school X this distribution was as

follows in the 1st-year class, using one of the well-known intelligence tests.

Score	*f*
95–99	1
90–94	5
85–89	8
80–84	17
75–79	13
70–74	24
65–69	25
60–64	35
55–59	29
50–54	24
45–49	9
40–44	11
35–39	5
30–34	4
25–29	0
20–24	1
15–19	1
	$N = 212$

$M = 63.02$
$\sigma = 14.30$

On the basis of this distribution the following limits for marks are assigned:

A	85–
B	71–84
C	56–70
D	42–55
E	–41

These "intelligence marks" should be transferred to the record card of each student. Then it would be the task of the registrar or secretary to send to each teacher each semester for each class a report that would look somewhat as follows:

First Semester 1926–27

French 1 *Teacher* A. M. Talbot

Your marks for this semester should correspond fairly closely with the percentages given below. Absolute correspondence is not expected. If you think more of your students deserve higher marks, you are at liberty to give them. If you feel that a larger proportion of your students should receive a non-passing mark, it is your privilege to give them. This statement is for your convenience to help you bring your marks in better accord with those of other instructors.

	Per Cent
A	14
B	26
C	46
D	10
E	4

By using a printed card all that would be necessary to fill in would be the semester, year, name of teacher, course, and per cents. The percentages would be obtained by making a frequency distribution of the intelligence marks of a class and changing these to per cents by means of a slide rule.

Note that the record of no individual student has been mentioned. The teacher is quite free to assign whom he will to *A*'s, *B*'s, *C*'s, *D*'s, and *E*'s. Quite possibly one of the students whose intelligence test placed him as one of the 14 per cent in the *A* group, might fall so low with reference to the others in the group in French, because of lack of interest, special disability, habits of study, or the like, as to warrant a *D* or *E* as a mark in French.

A scheme of this kind will also serve to inform the teacher as to his success in fitting his subject matter to student ability or preparation. If the teacher in French finds that his tendency is to mark too high, he should question whether he is recognizing initial ability or whether he is demanding a high enough rate of progress. On the other hand, if some teacher finds that his marks are running too low, he might question whether his students

were properly prepared when they came to him or whether his rate of progress is not too stiff for the class.

The use of test scores may be combined with this device for equating marks to give a still more valid system of marks. This might be done in the following manner:

1. Find the typical percentage distribution of marks for the whole school.
2. Find the σ limits for each mark by using a table of the probability integral.
3. Use this table of σ limits and apply to the distribution of intelligence scores and find the "intelligence mark" of each pupil.
4. Find the percentage distribution of "intelligence marks" for each recitation class.
5. Find the σ limits, using a table of the probability integral, for each of these "intelligence marks" in a given class.
6. Use these σ limits to assign marks to the class in terms of achievement test scores which have been given to the class.

It will be noted that no reference has been made to the percentage scale in any of this discussion. It is, therefore, entirely possible to assign letter marks without reference to the percentage system. That is, marks may be assigned according to the class distributions and not with reference to some assumed significance of the steps on the percentage scale.

One caution should be sounded. The method just described does not give any basis for determining who shall *fail* and who shall *pass*. In a strict sense, passing and failing has significance only in an absolute system of marking. When standardized tests become perfected, it will be possible to assign points on the scale of these tests which shall be the equivalent of passing at any stage of the course. This matter of promotion receives more extended discussion in the chapter devoted to that topic. But in a *relative*

system of marking (such as the system of letter marks) *there is no mechanical method of determining the passing point.* This cannot be emphasized too strongly. From superficial observation it would seem as though the method just described for turning test scores into marks automatically tells who should receive *F*'s and hence who should fail. On second thought it must be seen that the assignment of *D*'s and *F*'s by this method has no relation to passing or failing. *The determination of who shall pass and who shall fail in the traditional sense still depends on the judgment of the teacher, using all the evidence at hand.* In the illustration given there are 15 *D*'s and no *F*'s. But that does not necessarily mean that there should be no failures. It is quite possible that some or all of the 15 *D*'s should be failed. There is no way of determining this in lieu of absolute standards of passing except by using all the available data and the best judgment possible. The passing point on the percentage scale, 60%, 70%, 75%, or whatever it may be, looks like an absolute standard, but, as has been shown before in this book, it is an illusion.

An Index of Studiousness

A good measure of effort or of studiousness is much needed. The A R discussed in the chapter on "Derived Scores" fills that need in the elementary school, but, as has been shown, there are difficulties in the way of using the A R in the high school. The results of achievement and intelligence tests can be combined, however, in a simple way to give a measure of effort. Achievement test scores are the resultant of many factors. To simplify the matter we may say that achievement is the resultant of native ability being worked upon by various environmental forces. It is perhaps too much to say that intelligence tests measure native intelligence, although we know that such scores are highly correlated with native intelligence. If it were possible to eliminate from the achievement test scores the effect of intelligence test

scores, the result, according to the argument and within the limits of test reliability, would be the effect of environment which we may call *effort* or *studiousness*. Looking at the matter in another way, intelligence test scores give a measure of "average" ability, and predict what may be expected in the way of achievement. If more or less than the expected achievement results, the difference may be ascribed to greater or less effort or greater or less studiousness. If the reader would quarrel with the use of the words "effort" or "studiousness," let him substitute other words to describe the measure here proposed.

The elimination of intelligence may be accomplished in several ways. One method is to turn both test scores and intelligence test scores into ranks and to obtain the difference between the ranks. This difference between ranks is the measure of studiousness desired. Following is an example of the procedure on actual scores in a small class.

TABLE 127

STUDIOUSNESS INDEX, USING THE METHOD OF RANKS

Individual	Score on Achievement Test	Score on Intelligence Test	Rank on Achievement Test	Rank on Intelligence Test	Difference
A	22	130	13	14	1
B	24	132	11	13	2
C	26	154	9.5	8	− 1.5
D	22	144	13	12	− 1
E	38	185	1.5	2	.5
F	32	195	3.5	1	− 2.5
G	26	152	9.5	10	.5
H	22	171	13	5	− 8
I	29	182	7	4	− 3
J	38	162	1.5	7	5.5
K	30	153	5.5	9	3.5
L	32	170	3.5	6	2.5
M	28	183	8	3	− 5
N	30	148	5.5	11	5.5

Plus differences indicate good studiousness. Minus differences indicate poor studiousness.

A second method for obtaining this measure of studiousness is to obtain standard deviations and the differences between standard deviations. This latter method is shown on the same set of scores.

TABLE 128

STUDIOUSNESS INDEX, USING STANDARD DEVIATION UNITS

INDIVIDUAL	SCORE ON ACHIEVEMENT TEST	SCORE ON INTELLIGENCE TEST	S. D. ACHIEVEMENT	S. D. INTELLIGENCE	DIFFERENCE	INDEX OF STUDIOUSNESS
A	22	130	− 1.3	− 1.6	+ 0.3	53
B	24	132	− 0.9	− 1.5	+ 0.6	56
C	26	154	− 0.5	− 0.4	− 0.1	49
D	22	144	− 1.3	− 0.9	− 0.4	46
E	38	185	+ 1.9	+ 1.2	+ 0.7	57
F	32	195	+ 0.7	+ 1.7	− 1.0	40
G	26	152	− 0.5	− 0.5	0	50
H	22	171	− 1.3	+ 0.5	− 1.8	32
I	29	182	+ 0.1	+ 1.1	− 1.0	40
J	38	162	+ 1.9	+ 0.3	+ 1.6	66
K	30	153	+ 0.3	− 0.4	+ 0.7	57
L	32	170	+ 0.7	+ 0.4	+ 0.3	53
M	28	183	+ 0.1	+ 1.1	− 1.0	40
N	30	148	+ 0.3	− 0.7	+ 1.0	60

In this method the mean and standard deviation of the distribution of achievement scores and intelligence scores are found. Using these values (M = 28.5 and σ = 5.1 for the achievement scores; M = 161.5 and σ = 19.5 for the intelligence scores), an S. D. value is computed for each individual achievement score in column 4 and for each intelligence score in column 5. [To find the S. D. value for the achievement score of 22 of individual A subtract 22 from 28.5 (mean) giving 6.5. Divide 6.5 by 5.1 (S. D.), giving 1.3. The S. D. value is −1.3; minus, because the score 22 is below the mean 28.5.]

Column 6 is found by subtracting algebraically column 5 from column 4, being careful to give the correct signs. Since these difference values are confusing, some being plus and some minus, they are turned into more convenient units by moving the decimal point one place to the right and adding to or subtracting from 50 according to whether the difference value is plus or minus. These values are given in the seventh column and are called the index of studiousness. Average studiousness is represented by a value of 50. Values above 50 represent better than average studiousness and values below 50 less than average studiousness.

The results using the two methods are in agreement in the main. The first method using the difference between rank is the easier method and is preferred by some for that reason. Both methods show a regression effect. That is, it is extremely difficult, statistically, for a pupil with high intelligence to make a plus studiousness index, and vice versa. This is, indeed, impossible in the method using ranks. The pupil having a rank of 1 in intelligence can have no lower rank than 1 in achievement and hence the largest difference for this pupil is zero. Likewise the lowest possible achievement difference for the pupil lowest in intelligence is zero. This fault does not occur in the method using S. D. differences and the S. D. difference is the one recommended.

This measure of studiousness suffers from a fault common to most measures of conduct. It differentiates between the members of a single class but does not permit comparison between members of different classes. Only when one or more individuals are common members of two groups can a comparison between groups be made. Assumptions about equal average studiousness of different groups are fallacious, as so much depends on the motivation and the social setting. What is needed in this connection are standards on standardized tests which will permit the comparison between different groups. Such standards have been

determined for nine of the best achievement tests that measure high school subjects.[1]

Marking Sections of Different Levels of Ability

One of the most stubborn problems in marking has arisen in connection with the sectioning of pupils on the basis of ability. The difficulty is partly due to the unsatisfactory nature of letter marks which are assigned relative to other members of the group and not with respect to an absolute level of achievement. The difficulty is also due to the fact that marks are partly pedagogical in their function and partly administrative. The following diagram illustrates the dilemma. At the left is a diagram which represents a large group of about 100 pupils who distribute themselves in the normal curve in some school subject. The three

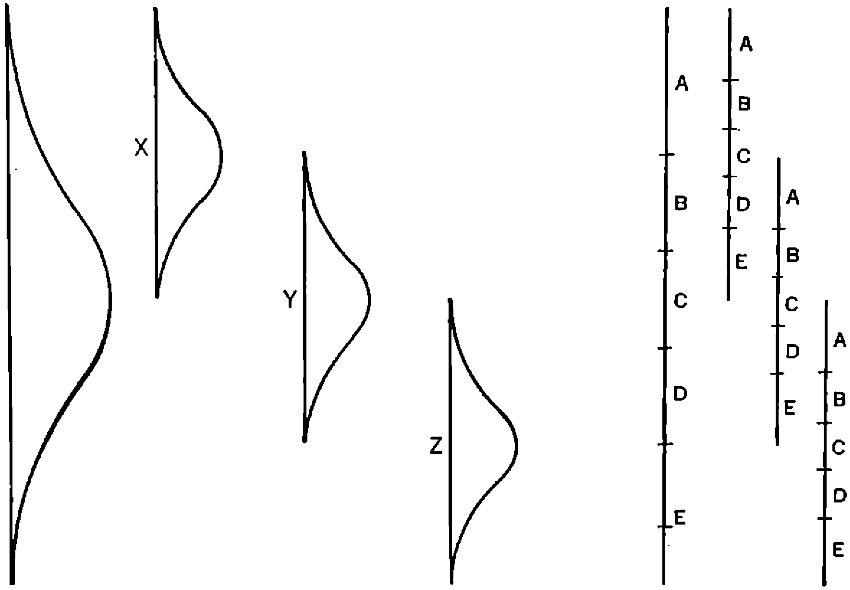


Figure 20. — Two methods of marking fast and slow moving sections

[1] Symonds, P. M. — *Ability Standards for Standardized Tests for the High School.* To be published by the Teachers College Bureau of Publications.

curves in the center represent this group of 100 pupils broken up into three groups on the basis of ability, labeled X, Y, Z, as is done in some cities to denote fast-, average-, and slow-moving sections.

There are two ways in which these sections might be marked, illustrated by the two scales at the right. They might be marked as though they still constituted part of the large group at the left. On this basis the X group would receive *A*'s, *B*'s, and *C*'s, with a majority of *B*'s. The Y group would receive a large preponderance of *C*'s with a few *B*'s and *D*'s. The Z group would average *D* with some *C*'s and practically all the failures. In assigning marks by plan No. 1 it is best to set department examinations, *i.e.*, coöperative, uniform examinations given to all of the sections of a class which can be scored and interpreted as though all the sections comprised one large class. The other plan would be to mark each section with the full range of marks from *A* to *F*, failing about the same number in each section. Plan No. 2 is the one which is largely used, mainly because it requires no breaking of old habits of marking on the part of teachers. In a large New York City high school I found that although freshmen were sectioned on the basis of ability, nothing had been done to alter the marking system. Some have suggested that plan No. 2 be used but the letters be given subscripts such as A_1, A_2, or A_3 to indicate whether the marks were given in a fast-moving, average, or slow-moving section.

Both plan No. 1 and plan No. 2 have their advantages, also their defects. From the administrative point of view plan No. 1 is to be preferred. The use of marks for promotion, college entrance credits, guidance, etc., would almost require that plan No. 1 be used so that the marks may have a definite meaning. If plan No. 2 is used, a mark has three different meanings according to the section in which it is given. On the other hand, plan No. 1 has a very bad effect on the two extreme sections. The pupils of

section Z, no member of which can ever, except under exceptional circumstances, get above a *B* or *C*, believe themselves branded as failures, and this leads to discouragement or despondency. The pupils of section X, finding that they can make *A*'s and *B*'s without effort, are deprived of a stimulus to effort. For pedagogical progress plan No. 2 is preferable. Of course, an absolute system of marks would solve the whole difficulty, for in that case marks would represent actual achievement. In the case of the two plans for giving letter marks, there seems no possibility of compromise—either plan No. 1 or plan No. 2 must be used. The only resolution of the dilemma is to give two marks according to each of the plans.

If administratively possible, the studiousness index should also be reported as a mark. The studiousness index carries a message which is not included in a mark of achievement alone, and which is a valuable supplement to the achievement mark. The studiousness index, for instance, should be the criterion for participation in athletics, as the reason for ineligibility in athletics is not so much to keep up certain levels of scholarship as to prevent pupils from neglecting their classroom mark. The studiousness index is a measure of the degree to which pupils attend to their classroom work.

In the lower grades there has been much discussion as to whether intelligence test scores, mental ages, or I Q's should be told to the children and to the parents. If pupils and parents were adequately informed as to the meaning of these terms and were intelligent enough to realize their significance and to take a sane attitude toward them, there would be no reason for not reporting them. Unfortunately it is not possible to make the significance of these items clear to parents and they are liable to misinterpretation. Even the Horace Mann School with a superior clientele does not give to the parents the results of its intelligence tests. In the case of high school pupils a principal is only creating trouble

for himself by reporting intelligence test results. However, if any particular parent or pupil asks for the results, he is entitled to them, with accompanying explanation, and pupils should be encouraged by their advisers to take an interest in their intelligence test scores and their significance.

In connection with the matter of marking sections of differing ability is the question as to whether the same distribution should hold for all the classes of the high school. It has been argued that since there is progressive elimination of the unfit in high school the level of marks should gradually rise throughout the years of the school. Starch presents data to show that this is actually the case and that there is such a rise in the typical school or college. There is a fallacy in this point of view. Those who maintain that the level of marks should be higher in senior than in freshman classes are unconsciously regarding the letter marking or percentage system as partaking of the nature of an absolute system which rises as it ascends to higher levels. But the letter-mark system is strictly a relative system. Relative to what? Relative to the whole school? If this is the case, freshmen should receive less than average marks. Usually each recitation class is considered as being a unit in marking. If this view is adopted the distribution of marks of seniors should be the same as the marks of freshmen. Another reason why it is believed that the marks should rise is that with the elimination of the unfit while they are freshmen there is progressively less elimination each year. There are two points of view here that are in error. One is that unless a boy is destined to finish high school he should be eliminated as early as possible in the high school course. To be sure, as high schools are now organized full value accrues only to the graduate. But schools should recognize that there is to be elimination and plan to benefit those who will leave before the end of four years as well as those who graduate. The other error in the point of view that marks should gradually be higher as the unfit are eliminated is that

the primary function of marks is to determine who shall pass and fail. As has already been pointed out, the determination of passing and failing is quite a different operation from that of assigning marks. Assigning marks is a form of measurement; determining passing or failing is a form of judgment, using marks as part of its data.

REFERENCES

CAMP, F. S. — "Some 'Marks': an Administrative Problem," *School Review*, 25 : 697–713, Dec., 1917.

CATTELL, J. McK. — "Examinations, Grades, and Credits," *Popular Science Monthly*, 66 : 367–378, Feb., 1905.

CORNING, G. B. — "The Meaning of Students' Marks," *School Review*, 24 : 196–202, March, 1916.

DEARBORN, W. F. — *School and University Grades. Ref.* Ch. 2.

FINKELSTEIN, I. E. — *The Marking System in Theory and Practice.* Warwick and York, Baltimore, 1913.

FLINNER, I. A. — "Rating Students on the Basis of Native Capacity and Accomplishment," *Educational Administration and Supervision*, 9: 87–98, Feb., 1923.

FOSTER, W. T. — "Scientific *versus* Personal Distribution of College Credits." *Ref.* Ch. 2.

GRAY, C. T. — *Variation in the Grades of High School Pupils. Ref.* Ch. 2.

HOLLEY, C. E. — *The Teacher's Technique*, ch. xiv. The Century Company, New York; 378 pp., 1922.

JOHNSON, F. W. — "A Study of High School Grades." *Ref.* Ch. 2.

KELLEY, T. L. — *Statistical Methods. Ref.* Ch. 12.

KELLY, F. J. — *Teachers' Marks. Ref.* Ch. 2.

Marking Slow-Moving Pupils. *Report of Committee of Secondary School Principals' Association, Los Angeles, California. Quarterly of Secondary Education*, 1 : 386–391, April, 1926.

MEYER, M. — "The Administration of College Grades," *School and Society*, 2 : 577–588, Oct. 23, 1915.

— "Experiences with the Grading System of the University of Missouri," *Science*, 33 : 661–667, April 28, 1911.

— "The Grading of Students." *Ref.* Ch. 2.

— "The Kind of Scholarship Records to Be Kept in Schools," *University of Missouri Bulletin 5, no. 32, Educational Series 8*, Nov., 1914.

MEYER, M. — "The Limit of Uniformity in the Grading of College Students by Different Teachers," *Science*, 40 : 530–532, Oct. 9, 1914.

ODELL, C. W. — "High School Marking System." *Ref.* ch. 1.

PICKELL, F. A. — "Credit for Quality in the Richmond (Ind.) High School," *Educational Administration and Supervision*, 3 : 537–538, Nov., 1917.

PRESSEY, S. L. — "Fundamental Misconceptions Involved in Current Marking System," *School and Society*, 21 : 736–738, June 20, 1925.

RUGG, H. O. — "Teachers' Marks and Marking Systems," *Educational Administration and Supervision*, 1 : 117–142, Feb., 1915.

— "Teachers' Marks and the Reconstruction of the Marking System," *Elementary School Journal*, 18 : 701–719, May, 1918.

STARCH, D. — "Can the Variability of Marks Be Reduced?" *School and Society*, 2 : 242 f., Aug. 14, 1915.

— *Educational Measurements*, ch., ii. The Macmillan Company, New York, 1916.

— *Educational Psychology*, ch. xxii. *Ref.* Ch. 2.

— "Reliability and Distribution of Grades," *Science*, 38 : 630–636, Oct. 31, 1913.

SYMONDS, P. M. — "Equating College Marks," *Educational Administration and Supervision*, 11 : 118–124, Feb., 1925.

— "An Index of Studiousness," *The Pedagogical Seminary and Journal of Genetic Psychology*, 32 : 257–265, June, 1925.

THORNDIKE, E. L., and BREGMAN, E. O. — "On the Form of Distribution of Intellect in the Ninth Grade," *Journal of Educational Research*, 10 : 271–278, Nov., 1924.

TIDYMAN, W. F. — "Adjusting Marking Systems to Differences in Groups," *School and Society*, 22 : 247–248, Aug. 22, 1925.

WELD, L. D. — "A Standard of Interpretation of Numerical Grades," *School Review*, 25 : 412–421, June, 1917.

WHITTEN, C. W. — "Standardizing and Making Uniform Teachers' Marks," being a *Committee Report Contained in the Sixth Yearbook of the National Association of Secondary School Principals*, 1922.

CHAPTER XXV

THE USE OF INFORMAL TESTS IN TEACHING

The new type tests described in Chapter III, besides serving as a form for standardized tests, may also be used in the construction of informal tests by the teacher. It will be long before standardized tests have been prepared for high school subjects so that the teacher can wholly dispense with tests of her own construction. It is the hope that this may be realized eventually to a large degree. The thread of argument of this book leads to the conclusion that sound measurement in the secondary school cannot be accomplished except through the agency of standardized tests. But in two senses standardized tests ought never wholly to replace the informal tests of the teacher's construction. In the first place no course in the high school ought to become so standardized and fixed and motionless that it can be caught and photographed by a standardized test. The high school curriculum ought always to be in a state of readjustment to keep up with the changing times, although this change need not be so rapid as now. Once the reorganization of the secondary curriculum is effected it will not be permitted to become static again. For this reason standardized tests can never wholly take the place of informal tests. From another point of view it would be fatal for teaching and the teacher to expect such a uniform curriculum that standardized tests could be used exclusively. The resourceful teacher requires some latitude in choosing material for the course, in shaping the course around local or personal interests, and in meeting the exigencies of local and contemporary events. Good teaching must strike when the iron is hot and this would not be possible with a rigid curriculum. There is no better way of making teaching artificial

and mechanical than to sharply define the course by examinations prepared from without. New York State has just such a system in the Regents' examinations. The Regents' examinations help maintain standards — this cannot be denied. Superior teachers can rise above the shackles of the Regents' examinations and can make their teaching vital and effective in spite of them, but the average teacher is reduced to a blind endeavor to make a good showing on the tests. The Regents' examinations are a mixture of bane and blessing.

Every teacher should study the form of new type tests and should be prepared to construct on these newer models informal objective tests for use in her own class. In making an informal test the teacher should survey the material to be tested and should note down the separate items of information or skills which it is desired to test. The number of these items depends upon the scope of the text. One should ask in an examination significant questions only on points which it is really important that the pupil remember. It is possible to dig out unimportant facts from a civics or history text that were never intended to be remembered but merely to illustrate more general propositions. Acceptable items should then be turned into true-false, multiple choice, single answer, completion or other forms. The true-false or single-answer types are the easiest to construct. The multiple-choice type is probably the best type, but it is often difficult to put all items into this form. It is very easy to make completion statements which are difficult to answer, but to make a satisfactory completion statement is probably most difficult of all.

In preparing the test the teacher may estimate the time from the following table taken from Ruch.[1] This time table is designed to get the simplest type of memory exercise. Items requiring thought require longer time.

[1] *Cf. The Improvement of the Written Examination*, p. 97. Quoted by permission of Scott, Foresman and Company.

TABLE 129

TIME TABLE FOR CONSTRUCTING NEW-TYPE TESTS FOR HIGH SCHOOL OR COLLEGE STUDENTS AND ADULTS (FROM RUCH)

Recall type	4 to 8 items per minute
Recognition type	6 to 10 items per minute
True-false type	10 to 15 items per minute

Ruch states that these rules are based upon the use of items of such difficulty that the average pupil will get about half of the items attempted correct, and the time limits are arranged so that about ninety per cent of all the pupils will have opportunity to attempt every item. He also adds that it is assumed that responses are to be indicated by the writing of a single word or short phrase, or by underscoring, or by some similar rapid device.

One advantage of the new-type test is that it may be immediately scored and discussed, thereby making the most of the discussion when interest in the test is running high. Usually when there is mutual respect between teacher and class, the teacher may allow pupils to exchange papers and correct the tests as the teacher reads the answers. The wrongs and omissions should be checked, the rights counted, and the score put on each paper before it is returned to its owner. Since the scoring is perfectly objective, little skill and no judgment is required on the part of the class to do this scoring. Indeed, questions of the new-type have been found so serviceable as a teaching device that *practice exercises*, using these forms, have been developed.

Class discussion should follow immediately. For this class discussion to be of greatest value, the test should be printed or mimeographed so that each child has the text of the statement or question under discussion. This precludes the use of the dictated true-false test. In the latter case the pupil has before him merely a column of pluses and zeroes and has to rely on his memory for separate questions. If discussion follows the test immediately, the reasons for each answer may be considered as

well as the rightness and wrongness of each response. The teacher can immediately correct false impressions that have hitherto escaped notice and can remedy weak spots in the teaching. Then again several statements are usually interpreted differently from the teacher's intention and these various interpretations can be caught immediately and adjusted, tending to leave the class in a better frame of mind if the discussion is skilfully handled. In Chapter III it was stated that no true-false or multiple-choice tests ought to be used for the measure of individuals unless there are at least 100 items. Rather than try to stretch a test out to this length over a limited amount of material, it would seem justifiable to combine the results of tests given on several separate days. One could give weekly true-false tests of 10 items each which would total 80 or 100 items at the end of the term. The writer personally prefers postponing this piecemeal checkup on study until the end of the term, when all 80 or 100 items can be given at once, but some teachers use the tests as a means of forcing study regularly throughout the term. On the other hand, there is an advantage in the test which requires a review and general reorganization of the work as a whole.

If several short tests are combined into a single test, there is the problem of weighting to be considered. If the score is to be made up as the summation of correct responses to many short items, probably not much injustice is done by counting each item as one point, even though the items differ considerably in validity and difficulty. This applies where one constructs a test having different types of questions, such as 25 true-false, 25 multiple choice, 20 matching, and 30 completion. Each of these items may count as *one*, and all the different parts may be added together. Or if daily or weekly tests have been given over a period of time, they also may be added together. Counting each question as 1 may not be exactly scientific and items of different sections of a test may be weighed inversely proportional to the variability of the

scores on that section, but this procedure is a refinement which is probably unnecessary. Pupils should not be judged on daily or weekly tests, nor on sections of a larger test which are so short as to have low reliability. Scores on short tests should not be turned into marks. The main reason for this is that short tests have relatively low reliability and it is bad practice to pass judgment on an individual when one is not sure of his judgment. Another reason for not turning the scores of a short test into marks is that one should not continually make the work of the school a mark-getting affair. The proper and normal interest in marks is encouraged if this interest is periodic and not continual. If the class demands an interpretation of the tests, the teacher may place the class distribution of marks on the blackboard and let each pupil see his standing relative to that of the other members of the class. This usually suffices as an interpretation of the test score.

Paterson [1] has much good advice about the larger organization of tests over a period of years. By using systematic methods one can economize much effort in the construction of new-type examinations and can construct better tests. Paterson suggests putting each test item on a filing card which may be filed away with information concerning the significance of the item. There are two items of information which should be obtained for questions on informal tests: the *difficulty* of the item and the *validity* of the item. Per cent of correct responses is a satisfactory measure of the difficulty of an item and this information should be noted on the filing card. To determine the validity the writer has employed the following procedure, which is a rough-and-ready method. Select the papers of the best twenty pupils in the class and the poorest twenty pupils in the class. This should not be done until the end of the term so that the best twenty and

[1] PATERSON, D. G. — *Preparation and Use of New-Type Examinations*, pp. 60–64, 1925.

poorest twenty may be determined by the term mark; to determine this from a single test is not reliable enough. Tabulate the pupils' correct responses on each item of the test. The measure of the goodness of the item is the difference between the number of correct responses for the good group and the poor group. The following is an example of the method:

TABLE 130

SHOWING METHOD OF DETERMINING GOODNESS OF TEST ITEM

ITEM NO.	CORRECT RESPONSES OF 20 BEST PUPILS IN CLASS	CORRECT RESPONSES OF 20 POOREST PUPILS IN CLASS	DIFFERENCE
1	18	11	+ 7
2	15	9	+ 6
3	12	12	0
4	8	10	− 2
5	7	5	+ 2

The best item in this example is the easiest item, — best because the difference 7 is the largest difference. There is little relationship between the difficulty of an item and the "goodness" of the item, although Vincent[1] has shown that the difficult item is more often a good item than is the easy item. Item 3 is a neutral item in the test — so much dead wood, as it fails to differentiate between the good and poor pupils in the class. Item 4 is positively "bad," inasmuch as more of the poor pupils than the good pupils answer it correctly. Using these methods, significant data may be accumulated concerning each test item and in the course of a few years one may have measures of validity and difficulty of several hundred items. The filing cards bearing these test items may be taken from the file at any time and conveniently used to type up into a test.

[1] VINCENT, E. L. — *A Study of Intelligence Test Elements*, TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 152, p. 34, 1924.

Informal tests may be made to yield valuable diagnostic information by tabulating the results of the test.

A teacher might have a blank form prepared on the mimeograph with the pupils' names typed in, which could be used for a weekly diagnosis, or as often as tests are given. A device like this would be most helpful in algebra, where frequent short tests are given. This tabulation sheet yields two kinds of information.

DIAGNOSTIC TABULATION SHEET

(Errors are checked; omissions are circled)

Pupil	Item No.													
	1	2	3	4	5	6	7	8	9	10	11	12	13	Total
A					X					X			o	3
B					X			X		X	X		o	5
C											X			1
D									X	X	X	X	X	5
E						o			o		o	X	o	5
F			X	X	X		X	X	X		X	X	X	9
G					X						X			2
H					X						X	X	X	4
Total			1	1	5	1	1	2	3	3	7	4	6	34

It tells which pupils are doing badly and which items were taught badly or learned poorly. We shall call these *pupil diagnosis* and *item diagnosis*, respectively. Pupil diagnosis tells at a glance which pupils are having the most and least difficulty in the class. Pupil C, for instance, made only one error and has practically mastered this particular topic. Pupil F, on the other hand, made 9 errors, which shows that he is careless and needs more drill before there is mastery. For those pupils having most difficulty and in need of remedial instruction one can see at a glance on exactly what items further help and what practice are needed. Item diagnosis informs the teacher which items are easiest and

hardest for the class. Items 1 and 2, for instance, were without error and the teacher may consider them well learned. Item 11, on the other hand, was missed by all but one in the class, which indicates either that it is a very difficult item or that the teaching of this particular item was defective. The teacher should follow this diagnosis by a review of her teaching procedure which should be analyzed, if possible, to determine where the teaching was weak. Reference to the test papers themselves may help reveal this weakness. The true-false test usually offers no clue to the type of error, but the other types of tests do yield some sort of evidence. Usually the teaching will be found defective in that this particular item was hurriedly taught, or some point was given undue or too little emphasis. Even though time is insufficient to reteach a topic, the teacher may use the evidence to make the teaching of this particular material better the next year.

In order to diagnose individual difficulties the test should contain several items of the same type so as to enable one to distinguish chance errors from recurring errors. A *chance* error is one that is made accidentally and not consistently. It is sometimes due to lapses of attention, to perseveration, or kindred slips. Recurring errors, on the other hand, are persistent errors that occur repeatedly when the same operation is required. Recurring errors are not accidental but represent wrong learning. In this sense the Rugg-Clark algebra tests are diagnostic by making use of a device of rhythmic repetition of test elements.

The "Briggs English Form Test" uses such a device. The test consists of 20 sentences with mechanical errors in the composition which must be corrected by the pupil. Seven different types of errors are represented.

Initial capital
Terminal period
Terminal question mark

Capital for proper noun or adjective
Run-on sentence
Apostrophe of possession
Comma before *but*

These are made to occur in exactly this order. Each error occurs five times. A tabulation sheet is provided similar to the form shown on p. 537.

The organization of the work of the school to take advantage of the knowledge furnished by diagnostic tests is a much-debated matter. Washburne argues effectively for individual instruction and progress at the pace of each individual. There is much to commend this system, such as the arousal of individual initiative in passing the hurdles provided, the provision for each pupil to advance at his own pace, the opportunity for the pupil to plan his own work and to plan his time to make most effective use of it. Many difficulties must be overcome in preparing material that is more or less self-teaching and that provides practice with opportunity for self-checking and in preparing tests which diagnose progress at frequent intervals. There are many advantages in the give and take in a class which are lost in individual instruction. The following discussion proposes a new organization of the work of the school to take advantage of individual differences and the possibilities opened up by diagnostic tests, which at the same time permits the retention of the class as a unit as at present.

In the first place, every topic ought, in reality, to be opened by a diagnostic test. In athletics, for instance, no instructor would attempt to instruct a group without finding out first what they can do. A swimming instructor first finds out whether any of the group can swim. If the majority can swim, then the instruction for a group is devoted to learning and perfecting new strokes, methods of breathing, etc. If the majority of the group cannot

swim, then the methods will be devoted to elementary strokes, to gaining confidence in and under water, etc. If the group is not homogeneous, those individuals with more or less than typical skill must wait until the instructor can give them individual instruction. But the typical classroom teacher starts in teaching the first day. The course of study is all mapped out — the teacher lays her plans before the opening of school. What difference if pupils in a geometry class know what a line, an angle, or 90 degrees is — all must memorize the formal definitions during the first week of the term. *The first act in the teaching process should be diagnostic testing in order that one may know what to teach.*

After teaching, what? Testing. And after *this* testing, what? Suppose that a physics teacher has finished the topic "magnetism" and has given test 13 in the Glenn-Oburn series. Suppose that the results are disappointing and that out of the 20 questions on the test only a handful are correctly answered on the average. The teacher does not feel justified in repeating the work on the topic. Usually the only recourse is to heave a sigh, to condole one's self that the class is so stupid, to resolve to do a better teaching job next year on that topic, to scold the class on its poor performance, and to let the pupils take the consequences in their final achievement in the course. The conscientious teacher may try to help individual pupils, but some arrangement should be made in the organization of the school for this remedial help.

The plan suggested here is to break up the class period into two parts, one part for work on advance topics and material for the whole class; the other part of the period to be devoted to remedial help with those pupils who show from a diagnostic test that they need such help. Only by some such arrangement can any advantage be gained from diagnostic testing. Otherwise a diagnostic test yields information, but nothing can be done about the condi-

tions revealed by the test. We have had a divided period to permit the organization of supervised study. The results are not very clear, but point in the direction that all depends on the skill and interest of the teacher. Effectively handled, supervised study yields valuable results, but so often it becomes a mechanical period in which pupils study without supervision, or in which supervision is crudely given and the teaching and recitation are allowed to linger on. The plan at present proposed requires a long period of perhaps an hour. During the first part of this period the advanced topic is discussed, the new material is developed and the assignment is made. Pupils who show satisfactory achievement on the preceding topic are permitted to spend the balance of the period as they see fit in study. Those pupils who showed unsatisfactory achievement on the preceding topic are given review or supplementary treatment. The exact technique of this remedial work will have to be worked out. It probably should not be a retreading of the same steps in the work. A new approach to the topic should be made if possible. Additional illustrative material should be used. If a demonstration is repeated, care should be taken to point out the critical points. New problem material should be employed. Everything possible should be done to make the work vivid and interesting. Such a divided period could be used to advantage in the languages, mathematics, and the natural sciences, that is, in subjects where the main objective is the acquisition of skill or information. In other classes, such as the social sciences, where the objective is the building of attitudes, or in English literature, where the main objective is appreciation, this divided period is unnecessary. Unless some such administrative device as the divided period is used it is difficult to see how full value can be obtained from diagnostic testing.

REFERENCES

BRIGGS, T. H. — "English Form Tests Alpha and Beta." *Teachers College Bulletin, no. 6, Fourteenth Series*, Dec., 1922.

PATERSON, D. G. — *Preparation and Use of New-Type Examinations.* World Book Company, Yonkers, 87 pp., New York, 1925.

RUCH, G. M. — *The Improvement of the Written Examination.* *Ref.* Ch. 3.

RUGG, H. O., and CLARK, J. R. — "Standardized Tests and the Improvement of Teaching in First-Year Algebra." *Ref.* Ch. 6.

RUSSELL, C. — *Classroom Tests.* Part II, particularly ch. ix, xiii, xiv, xvii. *Ref.* Ch. 3.

VINCENT, E. L. — *A Study of Intelligence Test Elements.* TEACHERS COLLEGE CONTRIBUTIONS TO EDUCATION, NO. 152. Bureau of Publications, Teachers College, Columbia University, New York, 44 pp., 1924.

APPENDIX A

STATISTICAL COMPUTATIONS

The Arithmetic Mean

In Chapter XII reference was made to a short method of computing the arithmetic mean. A detailed description of this computation is given below. The problem selected for treatment is the one already described on p. 231 and the description will start with the frequency distribution given in Table 65 on p. 233.

The first step is to guess the interval in which the mean lies. This is a pure guess. Simply guess as best you can where you think the average is. In a new column place a 0 opposite this class interval and number the intervals in each direction in regular order 1, 2, 3, 4, 5, etc. These are known as deviations. Deviations toward the higher scores should be marked plus and deviations toward the lower scores should be marked minus. Head this column *d*, standing for deviations.

Intelligence First-Year Boys

Scores	*f*	*d*
181–189	3	+ 5
172–180	1	+ 4
163–171	2	+ 3
154–162	4	+ 2
145–153	0	+ 1
136–144	2	0
127–135	3	− 1
118–126	2	− 2
109–117	1	− 3
100–108	3	− 4
91–99	1	− 5
82–90	2	− 6
73–81	1	− 7
64–72	1	− 8
	$26 = N = \Sigma f$	

The question is sometimes asked, "Where shall I guess to place the zero?" The answer is, "It does not make any difference where you guess. You will get the correct result no matter at which interval you assume the mean to be." As a matter of fact, the mean of this distribution really lies within the interval 127-135, but I have purposely guessed the mean one interval above that to demonstrate that it makes no difference where the mean is guessed to be. You will find, however, that your figures are smaller the more accurate you make your original guess.

The next step is that of multiplying the f column by the d column to get the fd column. For example, 3 times 5 is 15; 1 times 4 is 4; 2 times 3 is 6; etc.

INTELLIGENCE FIRST-YEAR BOYS

Scores	f	d	fd
181–189	3	+5	+15
172–180	1	+4	+4
163–171	2	+3	+6
154–162	4	+2	+8
145–153	0	+1	0
136–144	2	0	+33
127–135	3	−1	−3
118–126	2	−2	−4
109–117	1	−3	−3
100–108	3	−4	−12
91–99	1	−5	−5
82–90	2	−6	−12
73–81	1	−7	−7
64–72	1	−8	−8
$N = \Sigma f =$	26		−54

−54
+33
−21 add as in algebra = Σfd

$$26 \overline{)\,-21.000} = -.807 = \frac{\Sigma fd}{N}$$

20 8
200
182
18

−.807 = correction = c
9 size of interval
−7.263

140.0 mid-point of guessed interval
−7.3
132.7 = the Arithmetic Mean

After getting the fd column the steps are as follows:

(1) Find the sum of the fd column. The easiest way to do this is to add the plus numbers separately (33) and the minus numbers separately (−54) and add as in algebra. For those who have forgotten their algebra, the following reminders may be welcomed. The numbers 54 and 33 are really subtracted. Note that you must place the 54 above the 33 in order to subtract as usual

even though they do not occur in that order in the column. The answer, −21, has a minus sign because 54, the larger of the two numbers, is minus.

(2) Divide the sum of the *fd* column (−21) by *N* (26). Be sure to keep the minus sign. The result of this step is −.807.

(3) This quotient −.807 is multiplied by 9, the size of the class interval, giving −7.263. The size of the class interval is determined by subtracting 172 from 181, not 172 from 180.

(4) Subtract 7.3 from 140.0. 140.0 is the middle of the class interval in which you guessed the mean to be. There is an unfortunate difference in practice in the determination of the mid-point of the interval. Yule,[1] Kelley,[2] Otis,[3] Toops,[4] and others call 140 the mid-point of the interval 136–144. Garrett,[5] McCall,[6] Trabue,[7] Rugg,[8] and other educational statisticians call 140.5 the mid-point of an interval. Those that call 140.0 the mid-point of an interval intend that a score of 136 covers the interval 135.5 to 136.5, and the interval 136–144 covers 135.5–144.5. Those that call 140.5 the mid-point of the interval interpret a score of 136 as meaning 136.0–137.0 and the interval 136–144 as meaning 136.0–145.0. The assumption made by the latter group is that if a person makes a score of 136, he has made at least 136 and probably has tried some other item on the test which would make his score greater than 136 but less than 137. But these same writers are inconsistent. Suppose one were to find the average of a group of five scores 6, 8, 10, 12, 14. Taking the scores at their face value, one adds the scores (50) and divides by 5, giving 10. But the second group, according to their theory that a score of 6 really means 6.5, should add 6.5, 8.5, 10.5, 12.5, and 14.5 (52.5) with an average score of 10.5. There is always a difference of exactly .5 of a point between the answers from the two methods. The method advocated by Yule, Kelley, and others, taking scores at their face value (*i.e.*, 136 = 136.0), seems the more logical and is the one followed in this book. The mid-point of a class interval can always be simply determined by taking the average of the two extreme limits of the class interval; *i.e.*, $\frac{136 + 144}{2} = 140$. We sub-

[1] Yule, G. U. — *Introduction to the Theory of Statistics*, pp. 79–82, 112; 1919.

[2] Kelley, T. L. — *Statistical Method*, pp. 12, 13.

[3] Otis, A. S. — *Statistical Method in Educational Measurement*, pp. 50, 51.

[4] Toops, H. A. — "Computing Intercorrelations of Tests on the Adding Machine," *Journal of Applied Psychology*, 6: 172–184.

[5] Garrett, H. E. — *Statistics in Psychology and Education*, p. 7.

[6] McCall, W. A. — *How to Measure in Education*, pp. 367–368.

[7] Trabue, M. R. — *Measuring Results in Education*, pp. 145–154.

[8] Rugg, H. O. — *Statistical Methods Applied to Education*, pp. 85–87.

tract here instead of adding, because the sum of the *fd* column is negative. That indicates that we guessed our mean too high. Note further that in subtracting we have dropped off the last two figures in −7.263. We wish our answer correct to only one decimal place. We carry the quotient in step (2) to three decimal places in order to make sure that our work will be correct to the nearest figure in the first decimal at the end.

Those who are familiar with algebra may like to have the whole process summarized by means of a formula.

$$M = M' + \frac{\Sigma fd}{N} \times \text{class interval}$$

where M = arithmetic mean
M' = guessed mean
Σfd = sum of *fd* column
N = number of cases

TABLE 131

CHRONOLOGICAL AGES AND SCORES ON THE TERMAN GROUP TEST OF MENTAL ABILITY IN TWO CLASSES OF A SECONDARY SCHOOL

CHRONOLOGICAL AGE	FIRST-YEAR BOYS	TERMAN GROUP FORM B
16–7	J. McA.	64
14–9	H. M. A.	164
16–4	E. B.	144
15–11	W. B.	135
18–5	N. T. B.	80
14–4	G. W. DeV.	181
17–11	J. F. F.	101
15–10	M. B. F.	129
14–10	F. B. F.	158
13–4	R. E. G.	158
14–4	E. V. G.	176
15–0	K. B.	88
12–5	G. M.	122
14–0	E. B. M.	170
16–6	D. H. N.	136
14–2	J. P.	189
14–11	E. C. R.	187
15–2	D. F. R.	93
15–3	A. T. S.	159
14–3	A. H. S.	111

TABLE 131 — *Continued*

Chronological Age	First-Year Boys	Terman Group Form B
13–8	J. W. T.	161
14–5	A. B. V. H.	108
14–5	W. E. V.	132
17–2	T. W.	125
15–4	R. B. W.	108
13–6	E. R. W.	90

Chronological Age	Second-Year Boys	Terman Group Form B
14–11	W. A.	136
14–6	E. N. B.	165
16–4	L. S. B.	143
16–3	T. C.	141
16–9	H. M. D.	167
16–8	W. T. F.	184
16–1	W. C. F.	108
16–8	S. F.	122
15–2	F. G.	190
15–6	N. G.	140
13–6	K. G.	132
16–1	R. H.	138
16	A. K.	170
16–5	R. L.	117
14–9	R. L.	166
15–9	P. L.	190
15–2	F. McG.	168
15–2	W. M.	157
14–10	L. M. M.	180
17	C. N.	157
19–9	J. N.	135
15–1	O. S.	146
16–2	H. W.	139

The Standard Deviation

The same problem is continued to illustrate the computation of the standard deviation.

The new steps in this section for finding the standard deviation are as follows:

(1) Compute the column fd^2. $fd^2 = d \times fd$, and hence this column is found by multiplying the numbers in the d column by the num-

bers in the fd column. For instance, $5 \times 15 = 75$. Note that all the numbers in this column are positive, for when we get down to the minus deviations, minus times minus gives plus.

(2) Obtain the sum of the fd^2 column. (403)

(3) Divide the sum of the fd^2 column (403) by N (26), giving 15.50.

(4) Find the sum of the fd column as in the previous example. This is − 21, as before.

(5) Find c (correction) as in the last example by dividing the sum of fd (− 21) by N (26), giving − 0.807.

(6) Square c (− 0.81), giving +0.6561.

(7) Subtract c^2 (+.66) from the $\frac{\Sigma fd^2}{N}$, giving 14.84.

(8) Take the square root of the number obtained in step 7 (14.84), giving 3.85.

(9) Multiply by the class interval (9), giving 34.6, the S. D.

Scores	f	d	fd	fd^2
181–189	3	+ 5	+ 15	75
172–180	1	+ 4	+ 4	16
163–171	2	+ 3	+ 6	18
154–162	4	+ 2	+ 8	16
145–153	0	+ 1	0	0
136–144	2	0	+ 33	
127–135	3	− 1	− 3	3
118–126	2	− 2	− 4	8
109–117	1	− 3	− 3	9
100–108	3	− 4	− 12	48
91–99	1	− 5	− 5	25
82–90	2	− 6	− 12	72
73–81	1	− 7	− 7	49
69–72	1	− 8	− 8	64
$N = \Sigma f =$	26		− 54	$403 = \Sigma fd^2$

− 54
33
− 21

$-.807 = \frac{\Sigma fd}{N} = c$

26) − 21.000
20 8
200
182
18

$15.50 = \frac{\Sigma fd^2}{N}$

26)403.0
26
143
130
130

− .81
− .81
81
648
$+ .6561 = c^2$

14.84)3.85 $= \sqrt{\frac{\Sigma fd^2}{N} - c^2}$
9
68)584
8)544
765)4000
5)3825

15.50
.66
$14.84 = \frac{\Sigma fd^2}{N} - c^2$

3.85
9 class interval
34.65
34.6 = S. D.

The process that we have described has the additional step of subtracting the correction squared (c^2). This is because the deviations have been taken from the guessed mean and not from the true mean. If we had taken our deviations from the true mean, we would not have to make this correction, to be sure, but we would have much larger numbers to work with, each one containing decimal fractions, so that the method we have described is the shortest one for finding the standard deviation. The formula for finding standard deviation is

$$\text{S. D.} = \sqrt{\frac{\Sigma fd^2}{N} - c^2} \times \text{class interval}$$

The Coefficient of Correlation

THE PRODUCT MOMENT COEFFICIENT OF CORRELATION

The standard unit by which we measure the degree of correlation is known as the "product moment coefficient of correlation" or the "Pearson coefficient of correlation." Pearson was the first to give the correlation theory its precise mathematical formulation. The formula for computing the Pearson coefficient is $r = \frac{\Sigma xy}{N\sigma_x\sigma_y}$. x and y are deviations from the mean — the true mean — of each distribution. Σxy refers to the sum of the product of x and y deviations from their respective means. N is the number of individuals which are included in the distribution. σ_x and σ_y refer to the standard deviation of the two distributions. The following problem illustrates the use of the formula.

1	2	3	4	5	6	7	8
Individual	Score in X	Score in Y	Deviation from Mean for X d_x	d_x^2	Deviation from Mean for Y d_y	d_y^2	d_xd_y
A	1	8	− 2	4	− 1	1	+ 2
B	2	9	− 1	1	0		
C	3	7	0		− 2	4	
D	4	11	+ 1	1	+ 2	4	+ 2
E	5	10	+ 2	4	+ 1	1	+ 2
				$\Sigma d_x^2 = 10$		$\Sigma d_y^2 = 10$	$\Sigma xy = +6$

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{+6}{5 \cdot \sqrt{\frac{10}{5}}\sqrt{\frac{10}{5}}} = \frac{6}{10} = .60$$

In this example the mean of the X scores is 3. The deviation of 1 from this mean is -2. Other items in columns 4 and 6 are similarly obtained. The items in columns 5 and 7 are obtained by squaring the corresponding items in columns 4 and 6. The sum of column 8 gives Σxy. N in this example is 5. $\sigma_x = \sqrt{\frac{\Sigma x^2}{N}}$. Σx^2 is the sum of column 5 and Σy^2 is the sum of column 7. All of these values may be substituted in the formula and the computation leads to $r = .60$.

This problem is over-simplified. In any actual example the mean does not come out an integral value and consequently deviations from the mean would run into decimal values. In such cases we usually work from an assumed mean which we take as an integral value and then use a formula which makes a correction so that the true value of r is given.

The problem chosen to illustrate the computation of the product moment coefficient of correlation as it is usually found from the scatter-diagram is taken from data supplied by Mr. Luke C. Rhoads of the Mount Vernon, N. Y., High School. The results on the Hotz Algebra Test, Series *A*, Equation and Formula sheet, are correlated against the results in the Haggerty Reading Test, Sigma 3 Form A. Tables 132 and 133 are the work sheets. The steps in the work will be described in 1, 2, 3 order.

Step 1. Decide upon the *size* and *position* of the class intervals. One is working here with two distributions. The principles governing the making of distributions in general apply here. These principles have been briefly explained in Chapter 12, Section I. In the case of correlation the two distributions must be placed at right angles. Usually the limits of the class interval of one of the distributions is placed along the left of the scatter-diagram and the limits of the other along the top. In the example in Table 132 the scores along the left, for the reading test, have been grouped by fives. The scores along the top, for the algebra test, have not been grouped.

Step 2. Plot the points for each individual. For each individual there are two scores and these scores locate a square in the diagram. Let us suppose that the first individual whose scores are to be plotted has a score of 111 on the read ng test and of 1 on the Hotz test. One runs along the scale at the top of the sheet until he comes to the score "1." Then this individual lies somewhere in the column headed "1." Run down this column until opposite the interval 110–114 and place a tally. This definitely locates the score of this individual. This is done for all the individuals in the table. Finally the *frequency* of tallies in each square is transferred to Table 133.

TABLE 132

SCATTER-DIAGRAM: SCORES ON HOTZ EQUATION-FORMULA TEST AND SCORES IN THE HAGGERTY READING SIGMA 3

Score on Haggerty Reading Test Sigma 3	Score on Hotz Equation-Formula Test											
	0	1	2	3	4	5	6	7	8	9	10	11
130–134						II	I					
125–129					II							
120–124			I	I	I	II	卌					
115–119				I	III	III	卌			I		
110–114		I			IIII	卌	卌					
105–109	I			II	IIII	卌 I	III	I				
100–104	III		I	III	III	卌 III	IIII					
95–99			II	IIII	IIII	卌 卌	I					
90–94	II		I	I	IIII	卌	II					
85–89		II	IIII	I	卌	卌 III	I					
80–84	IIII	I		III	III	III						
75–79	I	II	II		II	II	I					
70–74	I	IIII			II	III						
65–69		III			I	II						
60–64		I	II			II						
55–59					I							
50–54												
45–49				I								
40–44	II											
35–39		I										

Step 3. Write the sum of the frequencies of each row at the right and the sum of the frequencies of each column at the bottom.

Step 4. Sum these columns, giving N. The distribution at the right is called the y distribution and its sum is known as Σf_y or N_y. The distribution along the base is called the x distribution nd its sum is known as Σf_x or N_x. $N_x = N_y = 191$, which is a check on the count.

Step 5. Assume a class-interval in both x and y distributions to be used as the guessed mean. In the y distribution this assumed mean is taken in the interval 95–99 and in the x distribution the assumed mean is taken as 4.

Step 6. Tabulate the unit-deviations in the column headed d for both deviations going in both directions + and − from the zero point which was taken for the assumed mean.

TABLE 133

The Computation of the Pearson Coefficient of Correlation

Score on Hotz Equation-Formula Test

Score of Haggerty Reading — Sigma 3	0	1	2	3	4	5	6	7	8	9	f	d	fd	fd²
130–134						2	1				3	7	21	147
125–129					4						4	6	24	144
120–124			1	1	1	2	5				10	5	50	250
115–119				1	3	3	5			1	13	4	52	208
110–114		1			4	5	5				15	3	45	135
105–109	1			2	4	6	3	1			17	2	34	68
100–104	3		1	3	3	8	4				22	1	22	22
													248	
95–99			2	4	4	10	1				21	0		
90–94	2		1	1	4	5	2				15	− 1	− 15	15
85–89		2	4	1	5	8	1				21	− 2	− 42	84
80–84	4	1		3	3	3					14	− 3	− 42	126
75–79	1	2	2		2	2	1				10	− 4	− 40	160
70–74	1	4			2	3					10	− 5	− 50	250
65–69		3			1	2					6	− 6	− 36	216
60–64		1	2			2					5	− 7	− 35	245
55–59					1						1	− 8	− 8	64
50–54											0	− 9	0	0
45–49				1							1	− 10	− 10	100
40–44	2										2	− 11	− 22	242
35–39		1									1	− 12	− 12	144
											191		− 312	2620
													− 64	
191	14	15	13	17	41	61	28	1	0	1	f			
	− 4	− 3	− 2	− 1	0	1	2	3	4	5	d			
− 144	− 56	− 45	− 26	− 17	125	61	56	3	0	5	fd			
− 19														
635	224	135	52	17		61	112	9	0	25	fd²			
	5	3	6	16		71	77	2		4	Σy +			
	45	72	31	22		79	8				Σy −			
							69	2		4	Σy +			
	40	69	25	6		8					Σy −			
587	160	207	50	6			138	6		20	Σxy +			
− 8						8					Σxy −			
579														

Step 7. Compute the *fd* column for both distributions by multiplying the corresponding members of the *f* and *d* columns.

Step 8. Sum the *fd* columns, taking care to conserve signs $\Sigma fd_{x'} = -19$ and $\Sigma fd_{y'} = -64$.

Step 9. Compute the fd^2 column for both distributions by multiplying the corresponding members of the *d* and *fd* columns.

Step 10. Compute the sum of the fd^2 columns. $\Sigma fd^2{}_{x'} = 635$ $\Sigma fd^2{}_{y'} = 2620$.

Step 11. The next operation is the computation of the sum of the products of the corresponding terms. This is done by working with deviations rather than with actual scores. The work is divided into three steps. In the first step the sum of the y' deviations is found for each separate deviation of the x''s. As this is usually done mentally, the $+ y'$ deviations and the $- y'$ deviations are found separately. In the example given, in the column $x' = 5$, there is only one individual whose value $y' = 4$. This value 4 is entered in the column Σy. There is no individual in the column $x' = 4$. In the column $x' = 3$ there is one individual whose value $y' = 2$ and 2 is entered in the column $\Sigma y'$. In the column $x' = 2$ there are several individuals whose y' deviations are summed by the following operations:

$$\begin{array}{cccccc} 1 \times 4 + & (2 \times 3) + & (3 \times 5) + & (4 \times 5) + & (5 \times 5) + & (7 \times 1). \\ 4 & 10 & 25 & 45 & 70 & 77 \end{array}$$

The first number in each parenthesis is the *y* deviation and the second number is the frequency of this deviation. Below the parenthesis is the cumulative total. All of this may be done mentally. There are also some negative y' deviations in the $x' = 2$ column, whose total is 8.

Step 12. The plus and minus $\Sigma y'$ values are combined by subtracting, leaving the difference in the column of the larger value. For instance, in the $x' = 2$ row there is a positive $\Sigma y'$ value of 77 and a negative $\Sigma y'$ value of 8. Subtracting gives 69 and it is placed in the plus column.

Step 13. Each value in the $\Sigma y'$ column is multiplied by the corresponding x' value, being careful to conserve signs. Thus $4 \times 5 = 20$, $2 \times 3 = 6$, $69 \times 2 = 138$, $-8 \times 1 = -8$, etc.

Step 14. The sum of the $x'y'$ column is found $\Sigma x'y' = 579$.

Step 15. These values are substituted in the formula to be used. The formula which yields the result with the simplest calculation is

$$r = \frac{\Sigma x'y' - \dfrac{\Sigma x' \Sigma y'}{N}}{\sqrt{\Sigma x'^2 - \dfrac{(\Sigma x')^2}{N}} \sqrt{\Sigma y'^2 - \dfrac{(\Sigma y')^2}{N}}}$$

The values found are substituted in this formula and the fraction is reduced to a decimal value.

$$r = \frac{579 - \frac{(-19)(-64)}{191}}{\sqrt{635 - \frac{(-19)^2}{191}}\ \sqrt{2620 - \frac{(-64)^2}{191}}}$$

$$= \frac{579 - 6.36}{\sqrt{635 - 1.89}\ \sqrt{2620 - 21.4}}$$

$$= \frac{572.64}{\sqrt{633.11}\ \sqrt{2598.6}}$$

$$= .447$$

In some cases one wishes to obtain the means and sigmas as well as the coefficient of correlation in the same computation. In this case the following formula is to be preferred:

$$r = \frac{\frac{\Sigma x'y'}{N} - \frac{\Sigma x'}{N}\ \frac{\Sigma y'}{N}}{\sqrt{\frac{\Sigma x'^2}{N} - \left(\frac{\Sigma x'}{N}\right)^2}\sqrt{\frac{\Sigma y'^2}{N} - \left(\frac{\Sigma y'}{N}\right)^2}}$$

where $$M_x = M'_x + \frac{\Sigma x'}{N}$$

$$M_y = M'_y + \frac{\Sigma y'}{N}$$

$$\sigma_x = \sqrt{\frac{\Sigma x'^2}{N} - \left(\frac{\Sigma x'}{N}\right)^2}$$

$$\sigma_y = \sqrt{\frac{\Sigma y'^2}{N} - \left(\frac{\Sigma y'}{N}\right)^2}$$

Note that all of the computation for the coefficient of correlation is done with deviations in terms of the unit of the class interval instead of actual score deviations. Since the coefficient of correlation is a ratio, there is no need to change the units into score deviations at any stage of the computation.

There are many cautions that one should note concerning the use of the product-moment coefficient of correlation. One that may be mentioned here is that the product-moment coefficient gives an interpretation of the degree of correlation only when the relationship is linear. This may be determined roughly by noting whether the plot of the scatter-diagram lies in a straight line. If the relationship is curvilinear, the product-moment coefficient of correlation gives values which are below the real relationship.

This computation, although complicated to describe, is comparatively simple after one is familiar with the procedure. It may be briefly summarized as the work of finding two standard deviations, the calculation of the sum of the product moments and the reduction of the formula in which these values are substituted. To carry through the computation from the above description may be confusing, but once the process has been carried out (with the aid of a teacher) several times, the steps in the process acquire meaning and the procedure becomes more straightforward and simple.

Various workers have prepared printed forms for plotting the correlation scatter and job analysis forms for facilitating computation.

Ruger has prepared a form which is merely a blank sheet of squared paper with the product moments, when the frequency of each square is one, printed in the corner of each square.

Toops has prepared a chart which uses a special device of summing along a diagonal which leads to a variation of the product-moment formula. Toop's chart is prepared so that the guessed mean is at zero for both of the variables in the plot. Ordinarily this would lead to very large numbers in the computation, but Toops uses a special device of grouping and then transmuting all original scores by means of a table, so that no score is over 18. Otis has derived a blank form which uses the same device of summing along the diagonal. Otis' chart uses deviations from an assumed mean which is taken as near as possible to the true mean.

Kelley, Thurstone, Holzinger, and Ruch have each devised charts which lead to computation by what is essentially the method described in this chapter. Kelley's chart offers a complete check on all the computation. Thurstone's chart is the one which would seem to be most satisfactory for all purposes. This chart is conveniently arranged and its computation procedure is as simple as any.

LIST OF PRINTED FORMS

Author	*Where Available*
Holzinger, K. J.	Author — University of Chicago, Chicago, Ill.
Kelley, T. L.	Stanford University Bookstore, Stanford, Cal.
Otis, A. S.	World Book Company, Yonkers, New York
Ruch-Stoddard	Extension Division — University of Iowa, Iowa City, Iowa
Ruger, H. A.	Teachers College Bureau of Publications, New York
Thurstone, L. L.	C. H. Stoelting Co., 3037 Carroll Ave., Chicago, Ill.
Toops, H. A.	Author — Ohio State University, Columbus, Ohio

II. THE RANK-DIFFERENCE FORMULA FOR COMPUTING THE COEFFICIENT OF CORRELATION

When one has ranked data, a much simpler method of computing a coefficient of correlation is available.

Rank in English	Rank in Algebra	Diff.	Diff.²
1	3	2	4
2	5	3	9
3	7	4	16
4	1	3	9
5	8	3	9
6	2	4	16
7	6	1	1
8	4	4	16
9	10	1	1
10	9	1	1
			82

The formula is $\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$.

D refers to the difference between corresponding ranks for an individual. Since these differences are to be squared, the signs may be ignored. N in the formula refers to the number of cases. The computation for the above example is as follows:

$$\rho = 1 - \frac{6 \times 82}{10 \times (100 - 1)}$$

$$\rho = 1 - \frac{492}{10 \times 99}$$

$$\rho = 1 - \frac{492}{990}$$

$$\rho = 1 - .496$$

$$\rho = .504$$

The symbol for this coefficient is ρ (rho) to indicate that it is a different coefficient from the Pearson product-moment coefficient. The Spearman rank-difference formula assumes a rectangular distribution. By means of a table (not given in this book) the value of the Spearman rank-difference formula may be transmuted into the Pearson product-moment formula. The difference between the two coefficients never amounts to as much as .02,

the correction always being added to the value of ρ. One would use the rank-difference formula only when the data are ranked, which is very seldom done when N is large. One may lay down a general rule that the rank-difference is not used when N is over 30. When N is under 30, the probable error of the coefficient is so large that the accuracy of the coefficient is to be mistrusted. The rank-difference formula, therefore, should only be used to indicate the existence of correlation and not to determine its exact amount.

Reliability of Measures

In this section formulæ are given for determining the reliability of measures.

RELIABILITY OF THE ARITHMETIC MEAN

No measure in education is absolutely accurate. A score obtained on any test, such as an intelligence test, is not necessarily the true score for the individual making it. Were the individual to repeat the test the probability is that he would not make the same score he made the first time. If we were to give the intelligence test to another sampling of boys of the same sort in the first-year class, the chances are that exactly the same arithmetic mean would not be obtained a second time. If we were able to repeat the test many times on children of the same sort, we would find that the arithmetic means would take many different values, all tending to cluster around some point somewhat as follows:

In Figure 21 there is represented in the large curve the distribution (smoothed) of a set of scores from a large group. Each dot along the base line represents a different position of the arithmetic mean as it would be

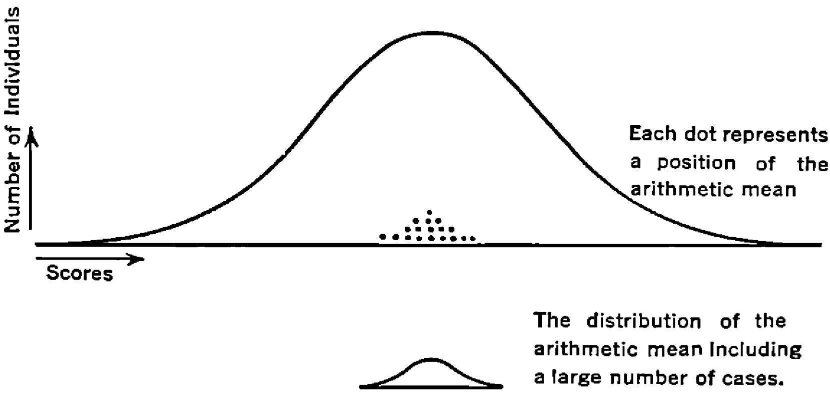


Figure 21. — Distribution of means

obtained in different trials — obtained from small groups selected out of the large distributions. In Figure 21 these dots are magnified and the curve that they tend to form is smoothed to represent the form that the curve would take, supposing that there were a large number of cases. This lower curve, the curve of the distribution of arithmetic means, were it possible to determine the mean a large number of times, happens to approximate very closely to the so-called probability curve which has already been mentioned. The mean of this (lower) curve we do not know from a single distribution of scores selected from the upper curve. That is, any dot or arithmetic mean in the upper curve may or may not be the average of the distribution of means (lower curve), which latter we take to be the true arithmetic mean of the total distribution of scores (upper curve). But we do know from a single distribution of scores the standard deviation of the distribution of means (lower curve). It is represented by the formula $\frac{\sigma}{\sqrt{N}}$, where σ is the standard deviation of the distribution of scores and N is the number of cases in the distribution. From this it may be seen that we know the mean more accurately the larger the number of cases in our distribution of scores, for in that case the above formula becomes smaller, which indicates that the variability of the distribution of means becomes smaller. In the problem in Chap. XII $\frac{\sigma}{\sqrt{N}} = \frac{34.6}{26} = 1.3$. This means that the distribution of dots in Figure 21 — the distribution of means — has a standard deviation of 1.3. From this we can say that we are very certain that the real mean of this first-year group of boys is within 3 times 1.3 of 132.7, that is, between 128.8 and 136.6; and the probabilities are that it is much nearer 132.7 than either of the extreme values we have mentioned. Three times the standard deviation is taken because, as has been already stated, that represents the practical limits of the curve.

To give one illustration of how we may expect the mean to vary in any group for different samples or selections I have divided the scores of the first-year boys into five lots picked at random. For the first lot I have chosen the 1st, 6th, 11th, 16th, and 21st boys; for the second lot I have chosen the 2nd, 7th, 12th, 17th, and 22nd boys, etc. This shows how the mean varies with different samples. These means vary with a larger standard deviation than for our whole distribution, because each of these lots has only 5 cases whereas our whole distribution has 26 cases. The larger number of cases tends to stabilize the mean.

1st Lot	2nd Lot	3rd Lot	4th Lot	5th Lot
64	164	144	135	80
181	101	129	158	158
176	88	122	170	136
189	187	93	159	111
161	108	132	125	108
5)771	5)648	5)620	5)747	5)593
154.2	129.6	124.0	149.4	118.6

RELIABILITY OF THE STANDARD DEVIATION

The S. D. itself also is a variable and we can express its standard error by the formula $\frac{\sigma}{\sqrt{2N}}$. That is, a standard deviation as found for any sample only approximates the true standard deviation of many similar samples, and should be expressed as $\sigma \pm \frac{\sigma}{\sqrt{2N}}$.

RELIABILITY OF THE PRODUCT-MOMENT COEFFICIENT OF CORRELATION

The reliability of the product-moment coefficient is expressed as a standard error in the following formula

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}$$

RELIABILITY OF THE RANK-DIFFERENCE COEFFICIENT OF CORRELATION

The reliability of the rank-difference formula is slightly larger and is equal to

$$\sigma = 1.05 \frac{1-\rho^2}{\sqrt{N}}.$$

The fact that our mean of 132.7 may vary for another sample of the same group between 128.8 and 136.6 is evidence that one ought not ascribe too great accuracy to the original determined value of the arithmetic mean. Any arithmetic mean found in educational data carried to two, three, or four decimal places is absurd. It would be absurd to write the mean as 132.137 when we are not even sure that the mean is 132 or not. Such figures carried to two or three decimal places lead the unwary to believe that we have

results that are highly accurate. A safe rule to follow is that given by Kelley:[1] "Keep to the place indicated by the first figure of one half the probable error."

To illustrate, take our problem

Mean = 132.737
Standard Deviation = 34.65
Population = 26

Probable error of the mean $\left(.6745 \frac{\sigma}{\sqrt{N}}\right) = 4.58$

Probable error of the standard deviation $\left(.6745 \frac{\sigma}{\sqrt{2N}}\right) = 3.24$

½ the probable error of the mean = 2.29
½ the probable error of the standard deviation = 1.62

Accordingly we publish the results as Mean = 133 and S. D. = 35.

This warning should not mean, however, that one should never *work* with no more than even digits. A general plan to follow is to work with one more decimal place than one finally intends to keep. For instance, if one intends to keep one decimal place, one should plan to work with two decimal places throughout. In discarding figures one should use care to express results to the nearest decimal place. For instance, 13.73 is nearer 13.7 than it is 13.8, so in expressing 13.73 with one decimal place, write 13.7. But 13.78 is nearer 13.8 than it is 13.7, hence 13.78 becomes 13.8. In the case of 13.75, which is midway between 13.7 and 13.8, write 13.8, choosing the 8 rather than the 7 because 8 is an even digit. Using this convention, one tends as often to raise as to lower his result in dropping figures and hence *avoids* a constant error of any kind.

[1] Kelley, T. L. — "How Many Figures Are Significant?" *Science*, 60: 524, Dec. 5, 1924.

APPENDIX B

TECHNICAL PROCEDURES IN DETERMINING TEST RELIABILITY

Methods of Determining Reliability

Reliability is measured by comparing the performance on two comparable tests. The most direct way of making this comparison is to find the coefficient of correlation between two comparable tests. This special coefficient of correlation is known as the reliability coefficient (symbolized r_{II}). If the reliability coefficient is + 1.00, the test is perfectly reliable, *i.e.*, a score is a perfectly accurate measure of an individual. If the reliability coefficient is .00, the test has zero reliability and no dependence whatever can be placed on test scores. Obviously the reliability of all tests is positive.

Where two comparable forms are not available it is customary to split the test into halves, and correlate these halves. The halves of the test may be considered as two forms of a test, each half as long as the original test. Hence the reliability coefficient is that for a test half as long as the original and must be corrected so that it may apply to the original. An example will illustrate the procedure. In the first place, in halving a test one is careful that the two parts are random selections from the whole test.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
×	×		×			×	×	×	×		×		×	×		×		×	×

Suppose that the above diagram represents each item in a test of twenty items. The checks represent the items correctly solved. The score is 13. It is customary in breaking the test up into random halves to take the odd-numbered items (1, 3, 5, 7, 9, etc.) for one half and the even-numbered items (2, 4, 6, 8, etc.) for the other half. In the above example the number of correct odd-numbered items is 6 and of correct even-numbered items is 7. Each paper is broken up into two parts in this way. The illustration used is that of individual E in the table.

Individual	Total Score	Odds	Evens
A	19	10	9
B	17	8	9
C	16	8	8
D	15	9	6
E	13	7	6
F	12	4	8
G	12	6	6
H	11	5	6
I	11	5	6
J	11	6	5
K	10	6	4
L	9	5	4
M	7	3	4
N	7	4	3
O	4	2	2

The correlation between the scores in the odd and even columns is .94. This coefficient is then corrected by a formula, known as the Spearman-Brown formula, which enables one to find the correlation when a test is doubled, trebled, or increased in length by any amount.

$$r_{af,\,Af} = \frac{a r_{1I}}{1 + (a - 1) r_{1I}}$$

$r_{af,\,Af}$ is read correlation between a forms of a test and A other comparable forms, a = number of forms totaled, r_{1I} = reliability coefficient of one form.

In our problem we may consider the reliability of each half the r_{1I} of the formula and we wish to find $r_{af,\,Af}$, where $a = 2$. The formula then reduces to (letting $a = 2$)

$$r_{1I} = \frac{2 r_{\frac{1}{2}\frac{I}{II}}}{1 + r_{\frac{1}{2}\frac{I}{II}}}$$

In our problem $r_{\frac{1}{2}\frac{I}{II}} = .94$. Substituting in the above formula

$$r_{1I} = .97$$

Thus when only one form is available, the reliability coefficient of the whole test is found by splitting the test into random halves, correlating the halves, and correcting by means of the Spearman-Brown formula for the whole test.

The Spearman-Brown formula also has been used to determine how much a test must be extended to secure a given reliability. Suppose that a test has a given reliability $r_{11} = .70$ and one wishes a reliability of .90 ($r_{af,\,Af} = .90$). These values are substituted in the formula and the resulting equation solved for a, the number of repetitions necessary.

$$.90 = \frac{a \times .70}{a + (a - 1)\,.70}$$

$$a = 3.9$$

That is, the test must be extended to 3.9 times its length to have the desired reliability.

There is one serious objection to the use of the reliability coefficient as a measure of reliability, namely, that it varies for the same test in different groups. Now a unit which is not constant leads to confusion. If one man claims the area of his property which is 12 is larger than the area of his neighbor's property which is 10, he should state his units to prove his case. If it should develop that his property is 12 acres and his neighbor's is 10 hectares (nearly 25 acres), his assertion is false. It is precisely the same with reliability coefficients. A test with reliability coefficient of .80 may actually be more reliable than another test with a reliability coefficient of .90. The reliability coefficient is dependent on the variability of the group in which it is found. The following formula is a statement of the relationship.

$$\frac{\sigma}{\Sigma} = \frac{\sqrt{1 - R_{11}}}{\sqrt{1 - r_{11}}}$$

σ and r_{11} refer to the variability and reliability as found in one group; Σ and R_{11} refer to the reliability coefficient as found in another group.

With the same test the reliability coefficient increases with the variability of the group. When reporting the reliability coefficient, one should always report the variability of the group in which testing was done. Reliability coefficients will automatically run higher if testing is done over a range of grades 7–12 in school than if the testing is confined to a single grade.

An entirely different means of expressing test reliability is to give the standard error or probable error of a score. The standard error of the

score of a test is a measure of the truth or accuracy or constancy of that score. Suppose that a pupil gets a score of 25 on a test. If, in giving a comparable form, one feels sure that the score will vary from 25 by only one or two points at the most, the standard error of the test is low and the test has high reliability. On the other hand, if one feels that the score may be 10 points or so higher or lower, the probable error of the test is large and the reliability low. I am indebted to Otis for the following illustration. Suppose a pupil takes 25 equivalent alternative forms of a 20-word spelling test and makes scores distributed as follows:

Score	*Frequency*	
14	1	
13	3	
12	2	
11	4	$M = 10.0$
10	5	$\sigma = 2.1$
9	3	
8	4	
7	2	
6	1	
	25	

The average score on these 25 tests is 10.0 and we may assume that this is the true score of the individual in question. But only 5 of the scores were actually 10 — they deviated on both sides of 10 and ranged from 6 to 14. The standard deviation of this distribution is 2.1 and this may be taken as the standard error of a score on this test. Multiplying by .6745, one has 1.4 as the probable error of a score. This is interpreted as usual that the chances are even that the true score lies within 1.4 points of an obtained score. Naturally one is never able to repeat a test 25 times on one pupil, so a different procedure is necessary in finding the standard error of a score. But the above illustration suffices, however, to indicate the meaning of standard error of a score.

In finding the standard error of a score it is customary to give one alternative form of the test and compute the difference in scores on the two forms, being careful to conserve signs. If the second score is larger than the first, label the difference positive; if the second score is smaller, label the difference negative. The following table shows these differences for test 1 of the "National Intelligence Test, Scale A," form 1, for 232 pupils.

Deviations	*f*	
+ 12	3	
10	2	
8	4	
6	16	$M = 1.1$
4	40	$\sigma = 3.4$
2	45	
0	65	
− 2	32	
− 4	14	
− 6	4	
− 8	4	
	232	

The mean of these deviations is the practice effect and should not be confused with the standard error of a score. The standard deviation of these deviations is 3.4. From this standard deviation the standard error of a score may be found directly by employing the formula

$$\sigma_{\text{score}} = \frac{\sigma_{\text{differences}}}{\sqrt{2}} = .707\ \sigma_{\text{differences}}$$

In this example $\sigma_{\text{score}} = .707 \times 3.4 = 2.4$
The probable error of a score may be found by the formula

$$\text{P. E.}_{\text{score}} = .6745 \times .707\ \sigma_{\text{differences}} = .4769\ \sigma_{\text{differences}}$$

which applied to our problem becomes

$$\text{P. E.}_{\text{score}} = 1.6$$

That is, the chances are even that the score of an individual in test 1 of form *A* of the National Intelligence Test differs by only 1.6 points from his true score on that test.

There is a mathematical relationship which connects this standard error of a score and the reliability coefficient expressed by the formula

$$\sigma_{\text{score}} = \sigma_{\text{distribution}} \sqrt{1 - r_{11}}$$

Knowing any two of these variables, it is possible to find a third. For instance, knowing the reliability coefficient of a test and the standard deviation of the distribution in which the reliability coefficient is determined, one can readily find the standard error of a score. For example, suppose one

finds the reliability coefficient of a test to be .85 in a distribution whose standard deviation is 14.5. Substituting in the formula,

$$\sigma_{score} = 14.5\sqrt{1 - .85}$$
$$= 5.6$$

This formula enables one easily to change a reliability coefficient into a standard error of a score and vice versa.

The formula $\sigma_{distribution}\sqrt{1 - r_{11}}$ is what is called mathematically invariant with respect to $\sigma_{distribution}$, that is, it is constant for any test, for distributions of any variability. This makes it preferable to the reliability coefficient as a measure of reliability because it is something that is always the same (within limits of sampling) for any test, regardless of the distribution in which r_{11} is found. But the σ_{score} has the disadvantage in comparing the reliabilities of two tests in that the unit of score of one test has no relationship to the unit of score in any other test. Two tests each with a standard error of measurement of 4 points of score may not necessarily be equal in reliability because 4 points on one test may not be equivalent to 4 points on another test.

Kelley [1] recommends that if the scores on two tests are equated in some way, one may use the standard error of a score as a satisfactory measure of reliability. $\frac{\sigma_{score}}{\sigma_{distribution}}$ would fulfill the above requirement and it is suggested that the above ratio may serve as a measure of reliability. Such a ratio has a definite relationship to the reliability coefficient, as shown in the following table.

[1] Kelley, T. L. — "The Reliability of Test Scores," *Journal of Educational Research*, 3: 370–379, May, 1921.

$\frac{\sigma_{score}}{\sigma_{distribution}}$	R_{II}
.00	1.00
.05	.9975
.10	.99
.15	.9775
.20	.96
.25	.9375
.30	.91
.35	.8775
.40	.84
.45	.7975
.50	.75
.55	.6975
.60	.64
.65	.5775
.70	.51
.75	.4375
.80	.36
.85	.2775
.90	.19
.95	.0975
1.00	.00

The conclusion of the matter is that there is no measure of reliability which is not in some way contingent. The reliability coefficient is contingent on the variability of the distribution in which it is found, and the standard error of a score is contingent on the size of the unit of the score of the test. Before two tests can be compared in reliability, a correction must be made in either the reliability coefficient or the standard error of the score.

APPENDIX C

LIST OF PUBLISHERS WHO PRINT TESTS

Tests in high school subjects are published by:

Bureau of Publications, Teachers College, Columbia University, New York City
Extension Division, University of Iowa, Iowa City, Iowa
Ginn and Company, Boston, Massachusetts
Gregg Publishing Company, New York City
Harvard University Press, Cambridge, Massachusetts
J. B. Lippincott Company, Philadelphia, Pa.
Public School Publishing Company, Bloomington, Ill.
Russell Sage Foundation, New York City
World Book Company, Yonders-on-Hudson, New York

These publishers will send their catalogues on request.

INDEX

ABBOTT, A., 109
Abbott and Trabue test of ability to judge poetry, 109
Ability, general, present, 485, 486
  grouping, 474–497
    administration of, 490–494
    advantages and disadvantages, 476–485
    basis of, 485–490
  sectioning. *See* Sectioning.
  standards, 329
  with Sentences Tests, 96–106
  See also Clerical ability, Levels of ability, Marking sections of different levels of ability, Mechanical ability, Mental ability.
Absolute marks. *See* Marks.
Absolute scale of intelligence, 302
Accidental error, 122
Accomplishment ratio, 265, 323–330
  causes of low or high, 328
  See also A. R.
Accuracy, 123
  in marking, 499, 500, 504, 507
  of test score, 288, 289
  of tests, 477, 478
Achievement, 326, 328, 364, 366, 374, 375, 400, 404, 466–468, 471–473, 484–487, 490, 491, 493, 498–500, 505, 506, 507, 510, 522, 524, 541
  defining and reporting, 22
  final, 489
  importance of correct judgment of, 20
  present, 489
  ratio. *See* A. R.
  test, 488, 489, 521, 525, 527
  See also Levels of achievement.
Acquaintance, effect on ratings, 348, 349
Administrative use of marks, 499–500
Admission to a course, 465
  to college, 411, 412, 438, 494
    requirements for, 438, 439
  to high school, 6, 390
Advice, 454, 456, 457
Age, 396, 401, 402, 407, 487
  norms. *See* Norms.
  See also Chronological age, Educational age, Mental age.
ALDRICH, T. B., 86
ALEGEYER, 381
ALEXANDER, H. B., 267
Algebra, objectives of instruction, 113
  prognosis of ability in, 365, 366
  standards in, 261–263
  tests, 114–125
ALLEN, R. D., 322, 323
ALLEN, W. S., 370, 373, 404, 405
Allen Latin prognosis tests, 370–372
ALLPORT, F. H., 338
Alternative forms of a test, 305, 306, 564
American Classical League, 148, 150
American Council Beta French and Spanish tests, 158
American Council intelligence test, 420, 426, 428, 441
American Council modern language tests, 156–161, 470, 472
  norms, 159
  Alpha tests, 158, 160, 502
Analogies test, 38, 436
ANGELL, 67
Anthropometric Committee of the British Association, 191
  measurement, 191–196

Aptitude, 380, 436, 454
  tests, 372, 377, 417
A. R., 323–329, 500, 521
  reliability of, 326
  tank conception of, 326
  use of as mark, 327
  use of as measure of teachers' efficiency, 327
Arithmetic computation, 261
  standards in, 543–560
Arithmetic mean, 234–236, 245, 254, 543–547, 554, 557, 558
  formula for, 546
  reliability of, 557–559
  standard error of, 558
Arkansas, University of, 434, 435
Army a, 67
Army Alpha, 56, 66–68, 73, 74, 267, 268, 380, 382, 383, 419, 420, 435
  equivalence of forms, 67
  norms, 68
Army Beta, 268
Army rating scale, 343
Artificial language test, 436
Ashbaugh, E. J., 262
Assumed mean, 550
Athletic proficiency, measurement of, 199–203
Attenuation of coefficients of correlation, 400
Attitude, school, 396, 400, 407, 417
Average, 254, 258, 259
  ability, 522
  deviation, 237
  error in rating, 351
Ayres, L. P., 95, 262
Ayres spelling scale, 95, 222, 280, 303

Bagley, W. C., 58, 325, 476, 477
Baldwin, B. T., 191, 192
Barr, A. S., 169
Barr diagnostic test in American history, 36, 169–173
  norms, 173
Barringer, 191
Beach, 213
Beatley, B., 417
Binet, A., 53, 54
Binet scale, 310, 311
Bingham, W. V. D., 56
Biology, objectives of instruction, 133
  tests, 133–135
Blackstone, E. G., 226
Blackstone tests of typewriting, 226–228
  norms, 228
  scoring, 227
Blood pressure measurement, 189, 190, 202, 203
Boas, F., 55, 191
Bobbitt, F., 184
Bowditch, H. P., 191
Bowers, Mrs. E. V., 97
Boyce, A. C., 340
Bregman, E. O., 511
Briggs, T. H., 95, 105, 370, 474, 480
Briggs analogies test, 372, 374
Briggs English form test, 105, 279, 305, 538
Brigham, C. C., 76, 268
Brightness, 321, 322
Brinkley, S. G., 26, 298, 299
Brooks, F. D., 312
Brown, A. W., 179
Brown, C. M., 220
Brown University, 422
Brown-Woody civics test, 179–181
  criticisms, 181, 182
  norms, 181
Buckingham, B. R., 265
Buckingham spelling scale, 95
Bureau of Public Personnel Administration, 381
Bureau of Public Personnel Administration clerical test, 381–383
Burg, C., 513
Burt, C., 326
Burt-Binet test, 477

C. A., 316, 320, 324
  See also Chronological age.
Cady, V. M., 355, 360
Caldwell, O. W., 493

CAMP, F. S., 473, 504, 505
CAMP, H. L., 135, 136, 138
Capacity, 328
Capitalization tests, 96–106
CAREY, G. L., 214–216
Carnegie Institute, 414, 417
Case School of Applied Science, 417, 420
CATTELL, J. McK., 54, 55, 510
C. E. E. B. examinations, 278, 292, 297
Censure, value of, 20
Central tendency, 237, 238, 245
CHAMBERS, O. R., 422
Chance error, 538
Changing sections, 491
CHAPMAN, J. C., 19
Character, 400, 401
ratings of, 400, 422, 507
traits, 421
CHARTERS, W. W., 105
Charters diagnostic language test, 105
Chemistry, objectives of instruction, 139
tests, 135–145
Chest capacity, measurement of, 195, 196
norms, 196
Chest girth, measurement of, 195, 196
Chicago, University of, 420
Chinning, 203
Choice, educational and vocational, 458, 469
of material, 279–285
CHOU, H. H. C., 91
Chronological age, 316, 319, 321, 396
See also Age.
CHU, J. P., 339
Circulatory system, 187
Civics tests, 179–181
CLARK, J. R., 120
Class interval, 234, 543, 545, 550, 554
Classes in a rating scale. *See* Rating scale.
Classical investigation, 148, 150, 156
Classical League. *See* American Classical League.
Classification, rules for, 468
test, 38
CLEM, O. M., 372, 401–404
Clerical ability, 379, 380
tests of, 377–383
COBB, M. V., 457
Coefficient of correlations, 245, 247, 252, 549–557
See also Job-analysis.
Coefficient of reliability. *See* Reliability coefficient.
Coefficient of validity. *See* Validity coefficient.
College entrance credits, 526
College Entrance Examination Board, 6, 19, 76, 428, 431, 433, 434, 435
See also under Examinations.
College entrance examinations, 415–418, 432, 437
College entrance requirements, 411, 412, 420, 421
College marks. *See* Marks.
College success, 416, 418, 421, 423, 427, 428, 441
See also Success.
COLLOTON, C., 319, 477
Columbia College, 412, 415, 426, 427, 428
Commission on the Reorganization of Secondary Education, 162, 221
Committee on Reorganization of Home Economics in Secondary Schools, 216
Committee on social studies, 162
Comparable tests, 288, 561
Competition, 498
Completion test, 37, 38, 54, 436
Composite, 367, 395, 401, 402, 515
of teachers' examinations, criteria of validity, 283, 284
of text books, criteria of validity 282

Composition ability, measurement of, 85–94
  scale, use of, 91
Comprehensive examinations, 417, 426
Computation, 543–560
"Conditions," 503–504
Conditions of pupils taking the test, 292–295
Conduct, measurement of, 333–361, 524
  objective tests of, 354–361
CONKLING, F. R., 97
Constancy of the I. Q. *See* I. Q.
Constant errors. *See* Errors.
Construction of objective tests, 25
Content examination, 437
COOK, R. R., 483, 490
COOK, W. A., 222
Correction, in finding mean and standard deviation, 548, 549, 550
  of coefficients of correlation for attenuation, 367
Correlation, 245–252, 349
  machine, 453, 454
  See also Multiple correlation, Partial correlation.
COSSMAN, L. H., 134, 283
COUNTS, G. S., 402, 449
Course of study, 491, 500
  trial, 451
COURTIS, S. A., 261, 262, 293, 374, 475
Courtis standard practice test in arithmetic, 261
COWDERY, K. M., 461
CRAMPTON, 191
Credit for graduation, 469, 494, 498
Criteria for the choice of tests, 278–309
Criterion, 279, 367, 379, 402, 426, 427
Criticism of standardized tests in algebra, 124
CROSS, E. A., 103
Cross English test, 103–105
Cross-out test, 38, 422
Curriculum, 412, 448, 449, 450, 458, 460, 500, 531, 532
CURTIS, F. D., 130, 133
Curve,
  of error, 239
  of mental growth, 317
  of normal distribution. *See* Normal distribution curve.
Curves, types of, 238, 245
Curvilinear relationship, 554
Cyclical test, 120, 121

DARBISHIRE, A. O., 247
Dartmouth College, 435
DEARBORN, W. F., 16, 321, 510
Decimal places, 559, 560
Democracy, 476, 477, 480
Demotion, 468
Department examinations, 526
Department of Superintendence of the N. E. A., 477
Derived scores, 310–332
Determinism, educational, 476, 477
Deviation, 237, 543, 554
  See also Average deviation, Root-mean-square deviation, Standard deviation.
Diagnosis, 16, 537, 538
  item, 537, 538
  pupil, 537, 538
Diagnostic test. *See* Test.
Diagnostic testing, 540, 541
Diagram, scatter. *See* Scatter.
Dickens' *A Tale of Two Cities*
  objective test on, 40
  key for objective test on, 49
Differences, 522–524
  S. D., 524
  sigma, 329
Difficulty of items of test, 535, 536
Directions, for giving a standardized test, 306, 307
  for giving true-false test, 29
  for scoring a test, 80, 81
Disciplinary value of marks, 505
Distribution, 236–239, 244, 247, 251
  limits of, 242, 243
  of A. R., 325
  of grade or age group, 511

Distribution — *Continued*
of I. Q.'s, 317, 318
of marks, 509, 528, 535
of means, 558
of scores, 232–234
See also Frequency distribution, Normal distribution.
Divided period, 541
Division of sections, 491
DODGE, C., 67
DOLL, E. A., 321
Drawing, tests of, 213–216
Duplicate forms, 305, 306
DVORAK, A., 133
Dvorak general science scales, 133
Dynamometer, 197
back, 197
Collin, 198
hand, 198
leg, 197
Smedley, 198

E. A., 324
Ease of scoring, 304, 305
EBBINGHAUS, H., 54
Economic condition of the home, 402
Educational age, 324
Effect of bunching on marks, 17
Efficiency, 8
Effort, 324, 400, 417, 522
ELDRIDGE, R. C., 222
Elementary school records, 396, 408
See also Marks, elementary school.
Elimination of pupils, 244, 528
ELLIOTT, E. C., 11
ELLIS, H., 55
EMERSON, W. R. P., 192
Emotional influence on pulse and blood pressure, 202
Energy, 396, 407
ENGELHARDT, J. L., 201
English, objectives of instruction, 84, 85
prognosis of ability in, 365, 366
tests, 84–112
E. Q., 324
Equating of marks, 517–521
Equation, regression. *See* Regression equation.
Error, accidental, 122
in scoring a composition, 89, 90
recurring, 121
variable, 352
See also Probable error, Standard error.
Errors, 538
chance, 538
constant, 347, 348, 350
criterion of validity, 281, 282
recurring, 538
Estimate, standard error of, 420
Estimates of intelligence, teachers', 396
Evenness of scaling, 290–292
Examinations,
advantages of new type, 39, 40
comprehensive, 417, 426
C. E. E. B., 292, 297, 298
content, 437
department, 526
fallacies in, 15, 16
final, 5, 506
limitation of new type, 40
old plan, 417, 432, 435
Regent's, 415, 423, 532
subject matter, 412, 415–418, 419, 420, 426, 431–433, 435, 441, 442
traditional, 3–5, 417, 418, 431, 441, 472
true-false, 25–37
See also Medical examination.
Extra-curricular activities, 498
Extrinsic use, criterion of validity, 280, 281

Failing, 465, 466, 468, 469, 472, 473, 520, 521, 529
Failures, 7, 235
Familiarity of pupils with the technique of taking tests, 295, 296
FARRAND, L., 54
Fast and slow moving sections, 525
FIFE, R. H., 156

Final examinations, 5, 506
status, 487, 488
FINKELSTEIN, I. E., 510
FLEMMING, C. W., 396, 400, 401, 404, 406
Foreign language tests, 148–161
Form for printing true-false examinations, 28
Formula for arithmetic mean, 546
product moment coefficient of correlation, 549, 553, 554
rank difference coefficient of correlation, 556
standard deviation, 549
standard error mean, 558
standard error of the product moment coefficient of correlation, 559
standard error standard deviation, 558
Formula of probable error of a score, 565
Spearman-Brown, 562, 563
standard error of a score, 565
standard error, rank difference coefficient of correlation, 559
FOSTER, W. C., 187
Foster test of physical efficiency, 187
FRANZEN, R. H., 259, 265, 268, 271, 324, 349, 506
FREEMAN, F. N., 58, 474, 477, 479, 481, 492, 494
French tests, 156–161
Frequency, 550
distribution, 233, 234, 240, 466, 519, 543
polygon, 240
FRETWELL, E. K., 393, 394, 398–400
FREYD, M., 346, 461
FROST, N., 266
FRYER, D., 454–456

GALTON, F., 55, 191, 337
GARRETT, H. E., 545
GATES, A. I., 245, 273, 374
Gates-Strang health knowledge test, 283
General ability, 485, 486
General science, objectives of instruction, 130
tests, 130–133
Geometry, demonstrative, 126
intuitive, objectives of instruction, 125
prognosis of ability in, 365, 366
tests, 125–129
German tests, 156–161
GERRY, H. L., 141, 142
Gerry test of high school chemistry, 141, 142
GILDERSLEEVE, 213
GLENN, E. R., 142
Glenn-Oburn physics tests, 142–144, 540
Glenn-Welton chemistry tests, 142, 144, 145
G. M. A., 311
Goal, 22, 235, 254, 260, 327, 466
GODDARD, 56
GODSEY, E., 155
Godsey diagnostic Latin comprehension test, 154
norms, 156
GOOCH, M., 429, 431, 439
Goodness of test item, 535, 536
GOWEN, G. W., 429, 431, 439
Grade norms. *See* Norms.
Graduation, 498
Graphic rating scale, 343–347
advantages of, 346, 347
rules to be followed in constructing, 346, 347
GRAY, 211
GRAY, P. L., 477
Gray oral reading test, 211
GREENBERG, J., 158
Gregg diagnostic shorthand tests, 226
Grip-norms, 197
Group, mental age, 311
test, 55
Groups, homogeneous. *See* Homogeneous grouping, Marking, homogeneous group.

Growth in mental ability. *See* Mental ability.
Guessing, on a true-false test, 29
  the mean, 544, 546
Guidance, 8, 322, 323, 373, 377, 385, 391, 448–464, 500, 526

HACKETT, 179
HAGGERTY, M. E., 106
Haggerty reading examination, sigma 3, 31, 106–108, 502, 550
  norms, 107
HAINES, 56
HALL, G. S., 360
Halo effect, 348
Handwriting, standards in, 262
HANLON, W. H., 268
HARPER, M. M., 478
Harvard University, 414, 416, 417, 428, 440, 493
HAUGHT, F. B., 339
Hawaii, University of, 414, 417, 423, 427
HAWKINS, H. E., 140
HAYES, 347
Health measurement, 192–196
Height-weight index, 192–195
Hemoglobin count, 202, 203
Henmon Latin tests, 484
High school marks. *See* Marks.
HILL, 226
HILLBRAND, 211
Hillbrand sight singing test, 211
HILLEGAS, M. B., 85
Hillegas scale for the measurement of quality in English composition for young people, 85, 338
History, objectives of instruction, 162, 163
  of intelligence testing, 53
  prognosis of ability in, 365, 366
  tests, 162–176
HOBBS, J. B., 272
HOKE, R. E., 222–225, 373
Hoke shorthand tests, 222–226
  norm, 224
  quality of writing, 224
  reading, 225, 226
  speed of writing, 224, 225
  vocabulary, 222–224
HOLLEY, C. E., 402
HOLLINGWORTH, H. L., 350–352
HOLMES, H. W., 493
HOLZINGER, K. J., 30, 555
Home, 402
  background, 400
Home economics information tests, 35, 217–219
Home Economics, objectives of, 216, 217
  tests of, 216–221
Homogeneous grouping, 245, 474, 481, 485, 487–489, 490, 491, 493, 494, 540
  See also under Marking
Honor marks, 421, 423
Honors, 498
Hop, step, and jump, 203
Horace Mann School, 396, 399, 407, 527
HOTZ, H. G., 114, 115
Hotz first-year algebra scales, 114–120, 124, 262, 263, 272, 279, 484, 502, 550
  manual for teachers, 115
  norms, 115, 116
  purposes, 124
  schedule of rotation, 115
  time for giving, 116
HUDELSON, E., 86, 91, 94, 95
Hudelson English composition ability scale, 86
Hudelson typical composition ability scale, 86
  directions for giving, 86
  norms, 88
  scoring, 88
HUGHES, W. H., 421, 422
HULL, C. L., 454
HURLOCK, E. B., 19, 20
HUTCHINSON, 213

I. E. R. general clerical tests, 378–380
I. E. R. general clerical test C-1, 379, 380

I. E. R. routine clerical test C-2, 379, 380
Improving tests, 25
Inadequacy of I. Q., 319–323
of mental age scale, 314–316
Incentives, 5, 20, 498, 499
Index. *See* Athletic index, Height-weight index, Schneider index, Sigma index, Studiousness index, Vital index.
Individual test, 55
Industry, 396, 400, 404, 407, 417
Inexpensive tests, 303, 304
Influence of illness on marks, 505
Informal test. *See* Test.
Informing parents of intelligence, 527
INGLIS, A., 108
Inglis vocabulary test, 108
Institute of Educational Research, 379, 385, 402
Intelligence, 53, 57, 318, 319, 326, 381, 385, 386, 396, 407, 439, 485, 500, 522, 524
absolute scale of, 302
first scale, 53
history of testing, 53
informing parents of pupils', 527
levels, 329
marks, 518, 519
quotient, 316–323
See also I. Q.
teachers' estimates of, 396
test, 374, 375, 400, 401, 405–407, 412, 452, 454, 459, 478, 484–490, 506, 517, 521, 527, 528
as predictive of college success, 418–426, 427, 428, 429, 435, 442
reliability of, 319
American Council intelligence test, 420, 426, 428, 441
Army Alpha, 66
Miller mental ability test, 71–73
Otis group intelligence scale, 58
Otis self-administering scale, 65, 315
Scholastic aptitude test, 76
Terman group test of mental ability, 68
Thorndike intelligence examination for high school graduates, 74
Thurstone psychological examination for college freshmen and high school seniors, 76
Interest, 460, 461, 481
questionnaires, 460, 461
Interpretation of a score, 259
Interval. *See* Class interval.
Invariant, 566
Iowa aptitude tests, 437
Iowa high school content examination, 434
Iowa physics tests, 135–138
criticisms, 137, 138
norms, 137
Iowa placement examinations, 417, 433, 436, 437
Iowa spelling scale, 95
I. Q., 257, 316–323, 324, 325, 328, 476, 477, 486–488, 490, 527
constancy of, 318, 319, 477
inadequacy of, 319–323
reliability of, 319
See also Intelligence quotient.
Italian tests, 56
Items, difficulty of, 535, 536
goodness of, 535, 536
number of, in informal test, 532, 533
selection of, 285, 286
test, selection of, Powers, 140
selection of, Gerry, 141, 142
validity of, 535, 536

JESSUP, K. E., 40
Job-analysis forms for computing coefficients of correlation, 555
JOHNSON, F. W., 16, 17, 507, 508
JOHNSON, H., 164
JOHNSTON, J. B., 421, 423, 424, 426, 440

JORDAN, A. M., 398, 404
Judgment, 333, 334
of experts, criteria of validity, 284, 285
Junior high school, 393

KEITH, J. A. H., 325
KELLEY, T. L., 95, 167, 284, 287, 288, 293, 364–366, 368, 392–394, 396, 399, 400–402, 406, 408, 452, 460, 545, 555, 560
KELLY, F. J., 16, 18, 507, 508
KEPNER, T., 176, 177
Kepner background tests in social science, 176–179
Key, scoring, 31
KIRBY, T. J., 19, 150
KIRK, J. G., 262
KLINE, L. W., 214–216
Kline-Carey drawing scale, 214–216
KNIGHT, F. B., 339, 348, 349, 355
KNOWLES, 222
KOOS, L. V., 162, 262
KWALWASSER, J., 207, 211, 212
Kwalwasser-Ruch test of musical accomplishment, 211–213
norms, 213

Language, (foreign) tests, 148–161
See also Artificial Language Test.
usage tests, 96–106
LA SALLE, J., 374
Latin, objectives of instruction, 148
prognosis of ability in, 370–372, 401, 402
tests, 148–156
Law of growth in a subject, 466, 467
Learning, 8, 9, 481–483, 489, 499
factors in, 19
rate of, 485–488
Length of test, 290
Letter marks, turning test scores into, 513–521
system of marking. *See* Marking systems.
Level of achievement, 525
See also Mental level.
Levels of ability, 329, 487–490, 517
See also Marking sections of different levels of ability.
LEWIS, E. E., 92
Lewis composition scales, 92
Limitations of absolute marks. *See* Marks, absolute.
of mental age norms, 258
Limits of distribution, 242, 243
of mental growth, 319
sigma, 516
Lincoln School of Teachers College, 493
LIPPMAN, W., 58
Load, pupil, 401
Lower quartile, 258

M. A., 310, 313, 316, 319, 324
See also Mental age.
MACPHAIL, A. H., 413, 415, 419, 422
MACQUARRIE, T. W., 388
MacQuarrie mechanical ability test, 388
Maine, University of, 429
Manometer, 197
elliptical, 197
Man-to-man scale, 341, 342, 343
Marking, homogeneous group, 494
letter, 5, 525, 527, 528
recitations, 506
relative, 520, 521
sections of different levels of ability, 525–529
standard of, 399, 400, 408, 414, 417, 418, 430, 431, 507–513
systems, 498–530
See also Percentage marking.
Marks, 5, 15, 391–393, 465, 469, 473, 498–530, 535
absolute, 21, 22, 500, 513, 520, 527
administrative uses of, 499–500
basis of, 504–507
college, 413–415, 417, 423, 426, 427, 429, 431, 441
reliability of, 431
disciplinary value, 505

Marks — *Continued*
effect of bunching, on, 17
elementary school, value of, 393, 394, 396
equating of, 517–521
function of, 498
high school, 397–399, 412–415, 416–420, 423–427, 429, 431, 432, 440–442
reliability of, 414
honor, 421, 423
influence of illness on, 505
letter, 20, 499, 502
turning test scores into, 513–521
limitations of absolute, 500
school, 400–402, 404, 405, 407, 408
reliability of, 401, 402
standardized, 22
term, 5
use of A. R. as, 327
variability of, 16, 17
working for, 499
MARSDEN, R. E., 477
Matching test, 36, 37
number of items, 36
Material, choice of, 279–285
Mathematics, prognosis of ability in, 366–370
tests, 113–124
Maturity, mental, 321
See also Mental growth, limits of.
MAY, M. A., 421, 423
MCCALL, W. A., 108, 260, 292, 293, 306, 379, 468, 545
MCKOWN, H. C., 438
Mean, arithmetic. *See* Arithmetic mean.
See also Assumed mean.
Measure of teacher's efficiency, use of A. R., 327
Measurement, injustice of imperfect, 19
uses of, 1, 531
See also Anthropometric measurement, Athletic proficiency, Blood pressure, Chest capacity, Chest girth measurement, Conduct, Health, Physical, Physiological measurement, Pulse measurement, Strength measurement.
Measures, reliability of, 557–560
Mechanical ability, 385, 386
tests of, 383–389
Median, 235, 258, 259
Medical examination, 201, 202
Mental ability, 321, 324
growth in, 312
age, 257, 258, 310–316, 321, 488, 527
See also M. A.
-age norms. *See* Norms.
-age scale, 302, 314
inadequacy of, 314, 316
growth, 312, 320, 321
limits of, 319
level, 321, 322
maturity, 321
MESSICH, 381
Methods of sectioning, 488, 489
MEYER, M., 510
Mid-points, 545
MILES, W. R., 16, 392
MILLER, W. S., 56, 57, 71
Miller mental ability test, 57, 71, 72 74, 258, 404
norms, 72
Minnesota, University of, 423, 424
MINNICK, J. H., 126
Minnick geometry tests, 126, 127
Missouri system of grading, 510
Mixed relations test, 38
Modern foreign language study, 158
tests, 156–161
MONROE, W. S., 109, 265, 297
Monroe standardized reading scale test III, 109
MORAN, 381
MOSHER, 213
Motivation, 271, 324, 329, 333, 499, 506, 524
Motive, 400
MOYER, E. L., 484

Multiple choice examination, 32
construction of, 32
giving, 33
number of statements, 33
scoring, 33
choice items, 286, 304
correlation, 366, 396, 401, 405, 452
track plan, 493
MURDOCH, K., 217, 219, 220
Murdoch sewing scale, 217, 219–221
norms, 220
Music, objectives in, 206, 207
tests, 206–213
Musical achievement, tests of, 201–213
capacity, 207–210

Nassau county supplement to the Hillegas scale, 86
National intelligence scales, 265, 273, 286, 293, 318, 319, 325–327, 374, 478, 484, 564, 565
New plan comprehensive examinations, 417
New type test, 473, 515, 531, 532
advantage of, 533
reliability of, 296–298
time-table for, 533, 534
use in teaching, 533, 534
New York City junior high schools, 471
NEWMAN, 192
Normal distribution, 242, 515, 516
curve, 239, 525, 557, 558
probability curve, 239, 240, 242, 490, 491, 510–512, 517
criticism of use of, in marking, 516
Norms, 254–259, 260, 261, 265–276, 302, 303, 459, 469
causes for variations in, 265–274
number of cases necessary in finding, 254
percentile, 258, 302, 303
types: grade, 254–256, 258
age, 255, 256, 258, 268, 314
mental age, 256–258
cautions in using, 257
limitations of, 258
variations due to difference in incentives, 270, 271
difference in locality, 266–268
difference in methods of teaching, 271–273
difference in race, 268, 269
difference in sex, 270
practice effect, 273, 274
American Council modern language tests, 159
Army Alpha, 68
Barr diagnostic tests in American history, 173
Blackstone tests of typewriting, 228
Brown-Woody civics test, 181
Chest capacity, 196
Godsey diagnostic Latin comprehension test, 156
Grip, 197
Haggerty reading examination, sigma 3, 107
Hillegas composition scale, 88
Hoke shorthand vocabulary test, 224
Hotz first-year algebra scales, 116
Iowa physics tests, 137
Miller mental ability test, 72, 73
Murdoch sewing scale, 220
Otis group intelligence scale, 63, 64
Otis self-administering test of mental ability, 66
Powers general chemistry test, 140
Pressey diagnostic tests in English composition, 103
Pressey-Richards test of understanding of American history, 175
Pressey test in Latin syntax, 156
Ruch-Popenoe general science test, 132
Rugg-Clark tests in first-year algebra, 122
Sanford-Schorling achievement test in plane geometry, 128
Schneider index, 190
Stenquist assembly test of mechanical ability, 385

Norms — *Continued*
Terman group test of mental ability, 70, 71
Thorndike intelligence examination for high school graduates, 75
Thurstone psychological examination for college freshmen and high school seniors, 76
Tyler-Pressey test in Latin verb forms, 156
Ullman-Kirby Latin comprehension test, 152
Van Wagenen American history scales, 270
Vital index, 196
White Latin test, 153
Number of sections. *See* Sections.
of statements in a true-false test, 27

Objective test. *See* Test.
Objectives: algebra, 113
biology, 133
chemistry, 139
English, 84, 85
general science, 130
geometry, intuitive, 125
history, 162, 163
home economics, 216, 217
Latin, 148
music, 206, 207
physical education, 184–187
physics, 135
stenography, 221
Objectivity, 289–290
of scoring, 25
of traits, 350–352
O'Brien, J. A., 271
Oburn, 142
Odell, C. W., 4, 502
Ohio State University, 417
Old plan examinations. *See* Examinations.
Omnibus test, 76, 284, 378
Opportunity, 328
Opposites test, 436
Order of statements in a true-false examination, 27
Orleans, J. S., 372
O'Rourke, L. J., 381
O'Shea, M. V., 222
Otis, A. S., 56, 58, 65, 256, 303, 304, 313, 545, 555
Otis group intelligence scale, 32, 38, 56, 58, 59, 73, 256–258
giving, 62
norms, 63, 64
scoring, 62
Otis self-administering test of mental ability, 65, 66, 74, 76, 311, 315, 396, 404
directions for giving, 65
norms, 66, 315
time allowance, 66
Overlapping, 244, 245, 471

Partial correlation, 364, 452
Partially standardized test for junior clerks, 381–383
See also Bureau of Public Personnel Administration Clerical Test.
Passing, 5, 6, 16, 259, 465, 466, 469, 472, 473, 520, 521, 529
Paterson, D. G., 347, 352, 535
Pearson, K., 55, 549–555
Pearson coefficient of correlation, 245, 549–555
See also Product moment coefficient of correlation.
Peckham, C. W., 191
Pennsylvania, University of, 54
Percentage marking, 4, 15, 16, 21, 499, 502, 504, 520, 528
Percentile, 259, 339, 424
norms. *See* Norms.
rank, 425
standing, 426, 440
Perrin, F. A. C., 435
Persistence in school, 396, 402
Personality traits, 421
Physical education,
aims, 184–187
directors, 186

Physical education — *Continued*
objectives, 184–187
tests, 184–205
See also Research.
Physical measurements, 186
Physics, objectives of instruction, 135
tests, 135–145
Physiological measurement, 187–191
PINTNER, R., 56
Pitfalls in constructing a true-false examination, 26
Placement, pupil, 465–469, 472, 477, 487, 499, 500
Polygon. *See* Frequency polygon.
POPENOE, H. T., 131
PORTER, W. G., 191
Posture, scale for measurement of, 187, 188
Power test, 117, 120
POWERS, S. R., 139, 140, 141
Powers general chemistry test, 34, 139–141, 287, 299, 502
norms, 140
Practice effect, 273, 274
Praise, value as incentive, 20
Predicted status, 485–487
Prediction, 7, 364–366, 372–375, 393, 394, 396, 460, 500
of success in college, 411–447
of success in high school, 390–410
of success in separate college courses, 429–437
Prerequisites, 244, 412, 439, 440, 465, 466
Present status, 485
PRESSEY, L. C., 155, 355
PRESSEY, S. L., 38, 97, 155, 173, 355, 422
Pressey diagnostic tests in English composition, 96–103
norms, 103
Pressey-Richards test of understanding of American history, 173–176
norms, 175
Pressey test of capitalization, 97, 98
Pressey test of grammar, 101–102
Pressey test in Latin syntax, 35, 155
norms, 156
Pressey test of punctuation, 98, 99
Pressey test of sentence structure, 99, 100
Pressey X-O tests, 422
Printed forms for plotting correlation scatters, 555
Prizes, 327, 328
Probability curve. *See* Normal probability curve.
integral, table of, 516
Probable error of a score, 563, 564
formula for, 565
Problems for research in measurement of physical education, 201–203
PROCTOR, W. M., 397
Product moment coefficient of correlation, 245, 549–555
formula for, 549, 553, 554
reliability of, 559
standard error of, 559
See also Pearson coefficient of correlation.
Proficiency. *See* Athletic proficiency.
Profile, 460
Prognosis, 7, 393, 436
of ability in algebra, 365, 366
English, 365, 366
geometry, 365, 366
history, 365, 366
Latin, 370, 371, 372, 401, 402
mathematics, 366–370
modern language, 373
stenography, 373
tests, 363–376, 406, 457, 458, 465, 466, 486–489
theory underlying construction of, 363, 364
Promotion, 6, 465, 473, 487, 498, 499, 526
Psychological examinations, 418–426, 436, 437, 441
See also Intelligence test.
Publishers of tests, 568
Pull-up, 198, 199
Pulse measurement, 189, 190, 202, 203

Punctuation tests, 96–106
Pupil load, 401
placement. *See* Placement.
selection, 244
Purposes of tests, 178, 179
Push-up, 198–199
PYLE, W. H., 55

Quartile, 258
Questionnaire, 360, 361, 461
reliability of, 360, 361
See also Interest questionnaires.
Quiz, 506

Range, semi-interquartile, 238
of scores, 244
Rank-difference formula of the coefficient of correlation, 556, 557
formula for, 556
reliability of, 559
standard error of, 559
Ranking, 338, 339, 340, 341
Ranks, 522, 524
percentile, 425
Rate of learning, 485–488
Rating, 333–354
methods, 187, 337–347
reliability of, 350–354
self, 349, 350
validity of, 347–350
variable errors in, 352
See also Trait.
Rating Scale, 344, 345
classes in a, 339, 340, 341, 346
See also Army rating scale, Graphic rating scale, Man-to-man scale.
Ratings, teachers', 393, 401, 404, 407, 408
RAUBENHEIMER, A. S., 355
Raubenheimer test of potential delinquency, 355–358
Reading tests, 106–109
true-false test aloud, 28
REASE, 381
Recitations, marking, 506
Record card, 500
Records. *See* Elementary school records.
Recurring error, 538
Regent's examinations. *See* Examinations.
Regression equation, 364, 452, 460
Re-grouping, 479
REJALL, 226
Relative marking systems, 520, 521
Reliability, 287–300, 341, 346, 351–353, 366
of A. R. *See* A. R.
coefficient, 326, 341, 350, 561, 565, 566, 567
factors which influence, 289–296
Gerry test of high school chemistry, 142
of high school marks. *See* Marks, high school.
of college marks. *See* Marks, college.
of intelligence test. *See* Intelligence test.
of the I. Q. *See* I. Q.
of measures, 557–560
of questionnaire. *See* Questionnaire.
of ratings. *See* Ratings.
repetitions necessary to get desired, 563
of tests, 477, 478
determination of, 561–567
what should one expect? 298–300
See also Arithmetic mean, New-type examinations, Product moment coefficient of correlation, Rank-difference coefficient of correlation, Standard deviation.
Remedial teaching, 540, 541
Repeaters, 469
Repeating class, 469, 470
Repetitions necessary to get desired reliability, 563
Report cards, 6
Reports to parents, 6
Requirements for college admission. *See* Admission to college.
in course of study, criterion of validity, 282, 283

Research, needed, in physical education, 201–203
needed, in prediction of high school success, 408–409
of college success, 442–443
Rewards, 328
Rho, 556
Rhoads, L. C., 550
Richards, R. C., 173
Right minus wrong method of scoring. *See* Scoring.
Rogers, A. L., 366–370, 402
Rogers, F. R., 196–199, 201
Rogers (sextet) tests of mathematical ability, 368–370
criticisms, 368, 369
Rogers test of mathematical ability, 305
Rollinson, E. A., 226
Root-mean-square deviation, 237
See also Standard deviation.
Ross, C. C., 394–396, 400, 402, 405, 406, 414
Ruch, G. M., 10, 15, 29, 30, 131, 134, 176, 211, 212, 282, 283, 284, 297–299, 532, 533, 555
Ruch-Cossman biology test, 34, 37, 134, 135, 283, 299, 502
norms, 135
Ruch-Popenoe general science test, 33, 131–133, 282, 476, 502, 514
norms, 132
Ruger, H. A., 555
Rugg, H. O., 120, 319, 342, 348, 353, 354, 477, 510, 545
Rugg-Clark tests in first year algebra, 117, 120–125, 538
norms, 122
purposes, 124
Ruggles, A. W., 380, 381
Ruggles test of clerical ability, 380, 381
Ruhlen, H., 97
Rules for constructing a true-false examination, 25
giving a test, 77–80
Sandiford, P., 274
Sanford, V., 127
Sargent, D. A., 200, 201
Sargent, L. W., 200
Sargent physical test of a man, 200–201
Scale, 21, 334
first, 53
units of, 22
values, 301
scoring by, 137, 138
See also Army rating scale, Mental age scale, Rating scale
Scaled tests, 300–302
Scaling, evenness of, 290–292
Scatter diagram, 424, 425, 550, 554
Schneider, E. C., 188, 189
Schneider index, 188–191
norms, 190
Scholastic aptitude test, 76
School marks. *See* Marks.
Schorling, R., 127
Schorling-Sanford achievement test in plane geometry, 127, 128, 502
norms, 128
Schwegler, R. A., 201
Science tests, 130–147
Score, interpretation of, 259
test. *See* Test.
T, 385
true, 564, 577
undistributed, 301
zero, 301
See also Derived score, Probable error, Standard error.
Scores, range of, 244
Scoring, 80, 81, 490
ease of, 304, 305
examination papers, 10
key, 31
objectivity of, 25
right minus wrong method, 29, 30, 32
scores less than zero, 32
a true-false test, 29–31
using scale values, 137, 138
Van Wagenen method, 167, 168

SCOTT, W. D., 67, 34?
S. D. Differences, 524
SEASHORE, C. E., 207–210
Seashore music tests, 207–210
SEAVER, J. W., 195
Sectioning, 7, 245, 322, 474, 475, 478, 479, 485–489, 493, 525
  methods of, 488, 489
  on basis of mental tests, 56
Sections, 475, 477, 483, 491, 492
  changing, 491
  division of, 491
  fast and slow moving, 525
  number of, 490
  size of, 490
  transfers of, 491, 492
Selection, 256
  of pupils, 244
  of test items. *See* Item.
Self-rating, 349, 350
Semi-interquartile range, 238
Sewing, tests of, 219, 221
SHEN, E., 349, 350, 352
SHORT, 381
Shorthand, tests of, 222–226
Sigma, 233, 237, 491
  difference, 329
  index, 259, 339
  limits, 516
Significant figures, 559, 560
Similarities test, 38
Sixteen spelling scales, 95, 96
  norms, 96
Size of sections, 490
Skipping, 491
SMEDLEY, F., 196
SMITH, D. E., 124
SMITH, S., 321
Snowball Fight on Slatter's Hill, 86–87
Social condition of the home, 402
Social science tests, 162–183
SOLOMON, M., 372
Spanish tests, 156–161
SPEARMAN, C., 556
Spearman-Brown formula, 353, 562, 563
  formula for, 562, 563
Spearman rank-difference formula, 556
Specific determinism in a true-false examination, 27
Speed, 123
  test, 117, 120
Spelling tests, 95–96
Speyer School, 399
Standard, 22, 465, 466, 469
  deviation, 236–238, 242, 245, 259, 523, 547–549, 554, 564, 565
    formula for, 549
    reliability of, 559
    standard error of, 558
  error, 327
    of estimate, 420
    of the mean, 558, formula for, 558
    of the product moment coefficient of correlation, 559, formula for, 559
    of the rank difference coefficient of correlation, 559
    of a score, 563–567, formula for, 565
    of the standard deviation, 558, formula for, 558
  of marking. *See* Marking.
Standardize, to, 275
Standardized, 257
  directions, 307
  mark, 22
  test. *See* Test.
Standards, 254, 257, 260–265, 280, 302, 303, 385, 500, 524
  ability, 329
  in algebra, determination of, 261–263
Standing, percentile. *See* Percentile standing.
Stanford achievement tests, 284
Stanford-Binet scale, 256, 257, 258, 310, 311, 314, 374, 477
Stanford University, 461
STARCH, D., 11, 16, 136, 475, 510
Statistical computation, 543–560
Status, final, 487, 488
  predicted, 485–487
  present, 485

STECHER, W. A., 199
Stecher, age norms of athletic performances, 199
Stencil, 305
Stenography, objectives, 221
  prognosis of ability in, 373
  tests in, 221
STENQUIST, J. L., 381, 383, 385, 386
Stenquist assembly test of mechanical ability, 383–386, 388
  norms, 385
Stenquist mechanical aptitude tests, 386–388
STETSON, F. L., 95
Stevenson Latin vocabulary test, 36
Stewartstown (Pennsylvania) High School, 514
STODDARD, G. D., 432, 434, 555
STONE, C. L., 435
STRANG, R., 283
Strength measurement, 197–199
STRONG, W. K., JR., 67, 461
Studiousness, 329, 507
  index of, 521–525, 527
Study, supervised, 541
  supervision of, 500
  time spent in, 404
Subjective test. *See* Test.
Success, school, 374, 391, 392, 395, 396, 400, 401, 405, 406–408, 459, 460, 498
  See also Research.
Supervised study, 541
Supervision, 541
  of study, 500
SYMONDS, P. M., 293, 340, 341, 414, 423, 426, 478

Tabulating, 232–234
Tank conception of A. R., 326
Teachers College, 493
Teacher's efficiency, use of A. R. as measure of, 327
  estimate of intelligence, 396
Teaching, 531–542
TEAGARTEN, F. M., 320
Term mark, 5
TERMAN, L. M., 56, 68, 201, 202, 257, 258, 265, 285, 286, 317, 318, 320, 439, 457, 493
Terman group test of mental ability, 39, 56, 68, 69, 70, 71, 73, 231, 235, 254, 256, 258, 259, 286, 316, 322, 396, 404, 420
  norms, 70, 71
Test, 361, 506
  ability with sentences, 96–106
  accuracy of, 477, 478
  algebra, 113–125
  analogies, 38, 436
  aptitude, 372, 377, 417
  biology, 133–135
  capitalization, 96–106
  chemistry, 135–145
  civics, 179–181
  classification, 38
  clerical ability, 377–383
  comparable, 288, 561
  completion, 37, 38, 54, 436
  conduct, objective, 354–361
  construction of objective, 25
  criteria for the choice of, 278–309
  cross-out, (X–O), 38
  cyclical, 120, 121
  diagnostic, 121, 539, 540
  directions for scoring, 80, 81
  drawing, 213–216
  English, 84–112
  foreign language, 148–161
  French, 156–161
  general science, 130–133
  geometry, 125–129
  German, 156–161
  group, 55
  history, 162–176
  home economics, 216–221
  improving tests, 25
  individual, 55
  inexpensive, 303, 304
  informal, 5, 473, 507, 515, 531–542
    making an, 532
  Italian, 156
  language usage, 96–106

Test — *Continued*
Latin, 148–156
length of, 290
matching, 36, 37
mathematics, 113–129
mechanical ability, 383–389
mixed relations, 38
modern foreign language, 156–161
See also Artificial language test.
multiple choice, 32–35
music, 206–213
musical achievement, 211–213
musical capacity, 207–210
objective, 417, 418, 435, 437, 441, 507
sample of, 40
omnibus, 76, 284, 378
opposites, 436
physical education, 184–205
physics, 135–145
Power, 117, 120
Prognosis tests. *See* Prognosis.
publishers of, 568
punctuation, 96–106
purpose of, 178, 179
reading, 106–109
rules for giving, 77–80
scaled, 300–302
science, 130–147
scores, accuracy of, 288, 289
selection of items, Powers, 140
Gerry, 141, 142
sewing, 219–221
shorthand, 222–226
similarities, 38
social science, 162–183
Spanish, 156–161
speed, 117, 120
spelling, 25, 95–96
standardized, 22, 62, 85, 276, 278, 279, 280, 285, 429, 466, 472, 473, 500, 507, 508, 510, 515, 521, 524, 531
directions for giving, 306, 307
stenography, 221
subjective, 289, 432, 433
time for administering, 490
timing of, 79, 123
true-false, 25–32
typewriting, 226–228
vocabulary, 106–109
weighting of. *See* Weighting.
yes-no, 31
See also Intelligence test, Reliability.
THEISEN, W. W., 484, 485
THORNDIKE, E. L., 54, 55, 57, 58, 67, 108, 117, 123, 213, 214, 215, 261, 262, 274, 300, 302, 312, 317, 320, 321, 338, 348, 354, 355, 360, 374, 378, 379, 381, 428, 441, 451, 452, 475, 511
Thorndike drawing scale, 213, 215
Thorndike handwriting scale, 338
Thorndike intelligence examination for high school graduates, 57, 74–76, 274, 419, 420, 423, 426, 428, 435, 441
norms, 75
Thorndike-McCall reading scales, 108, 151, 260, 273, 326, 396
Thorndike test of word knowledge, 108
THURSTONE, L. L., 57, 76, 228, 377, 378, 381, 420, 429, 436, 555
Thurstone clerical test, 377, 378
Thurstone examination in typing, 228
Thurstone psychological examination for college freshmen and high school seniors, 57, 76, 414
norms, 76
Time for administering tests, 490
spent in study, 404
-table for new type tests, 532, 533
Timing of a test, 79, 123
TOOPS, H. A., 297, 298, 379, 381, 385, 386, 418, 419, 420, 452, 545, 555
TORGESON, 213
TRABUE, M. R., 109, 211, 545
Tracks, 491
Traditional examinations. *See* Examinations.

Trait, 334–336, 349, 352
objectivity of, 350–352
personality, 421
rating, 421, 422
Transfers of sections, 491, 492
Transmute, 339
Trial course, 451
True-false examination, 25
directions for giving, 29
form for printing, 28
guessing, 29
number of statements, 27
order of statements, 27
pitfalls in constructing, 26
reading aloud, 28
rules for constructing, 26
scoring, 29–31
"specific determiners," 27
True score, 564, 577
Tryon, R. M., 163
T-score, 385
Turning test scores into letter marks, 513–521
Tyler, C., 155
Tyler-Pressey test in Latin verb forms, 155, 156
norms, 156
Typewriting, tests in, 226–228

Ullman, B. L., 150
Ullman-Kirby Latin comprehension test, 150–152
norms, 152
Undistributed score, 301
Units, 420, 421, 439
of scale, 22
University. *See* under name of university.
Upper quartile, 258
Upton, C. B., 115
U. S. Bureau of Education, 448

Validity, 279–286
coefficient, 419, 420, 427
of items of test, 535, 536
of ratings, 347–350
Ruch-Popenoe general science test, 132
Van Denburg, J. K., 402
Vannest, 179
Van Wagenen, M. J., 92, 94, 109, 133, 140, 165, 167, 168, 270
Van Wagenen American history scales, 165–169, 502
character judgment scales, 165
information scales, 165
thought scales, 165
Van Wagenen composition scales, 92
Variable errors in rating, 352
Variability, 237, 238, 244, 245, 466, 467, 563
of marks, 16, 17
Variations in norms. *See* Norms.
Varying work to fit ability, 493, 494
Vincent, E. L., 536
Vital index, norms, 196
Vocabulary tests, 106–109
Vocational advice, based on best evidence obtainable, 19
Voelker, P. F., 350, 354, 355

Washburne, C. W., 486, 539
Watson, J. B., 67
Webb, E., 348, 350
Weidemann, C. C., 26–30
Weighting of tests, 364, 367, 381, 402, 515, 534
Weld, L. D., 509
Wells, G. C., 40
Wells, 56
Welton, 142
West, G. M., 191
Whipple, G. M., 56, 71, 196, 198
White, D. S., 152
White Latin test, 34, 152–154, 502
norms, 153
Whitman, A. D., 418, 431, 432, 434
Whitmire, E. D., 265
Wilkins, 373
Williams, J. F., 184
Wilson, G. M., 105, 281
Wilson language error test, 105, 281
Wood, B. D., 13, 14, 19, 30, 158, 292, 297, 306, 415, 423, 426, 427, 428, 471

WOOD, T. D., 192
WOODS, F. A., 55
WOODWORTH, R. S., 67, 355, 381
Woodworth-Cady questionnaire, 359, 360
WOODY, C., 179
Woody-McCall mixed fundamentals arithmetic scale, 325, 327, 396
WOODYARD, E., 95, 295, 478
Working for marks, 499
World War, 56
WUNDT, 54, 349

YERKES, R. M., 56, 67
Yes-no test, 31
  rules for, 31
YOAKUM, C. S., 67
YULE, G. U., 545

Zero score, 301